

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Matej Žniderič

Spletni iskalnik podatkov o osebah

DIPLOMSKO DELO

UNIVERZITETNI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

MENTOR: doc. dr. Dejan Lavbič

Ljubljana, 2013

Rezultati diplomskega dela so intelektualna lastnina Fakultete za računalništvo in informatiko Univerze v Ljubljani. Za objavlanje ali izkoriščanje rezultatov diplomskega dela je potrebno pisno soglasje Fakultete za računalništvo in informatiko ter mentorja.



Št. naloge: 00148/2013

Datum: 10.09.2013

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **MATEJ ŽNIDERIČ**

Naslov: **SPLETNI ISKALNIK PODATKOV O OSEBAH**
WEB SEARCH ENGINE OF PERSON DATA

Vrsta naloge: Diplomsko delo univerzitetnega študija prve stopnje

Tematika naloge:

Na svetovnem spletu obstajajo številni podatkovni viri, ki vsebujejo podatke o fizičnih osebah, vendar med njimi ne obstaja semantična povezava. Potreba za takšen integriran in povezan vir podatkov zagotovo obstaja na številnih področjih. Če je na voljo odprt dostop do podatkov, kot je to urejeno v nekaterih državah, je seveda takšno povezovanje zelo poenostavljeno, v nasprotnem primeru pa je potrebno uporabiti druge pristope luščenja podatkov. V okviru diplomske naloge naj študent identificira zbirke podatkov o fizičnih osebah v Sloveniji in pripravi pregled, katere podatke takšni viri hranijo. V okviru implementacije naj pripravi delujoči prototip aplikacije, ki vključuje nekaj izbranih virov z osebnimi podatki fizičnih oseb in prikaže, kako lahko z integracijo različnih virov dosežemo učinkovitejše iskanje pod podatkih o ljudeh in predvsem pridobimo celovitejši pogled.

Mentor:

doc. dr. Dejan Lavbič



Dekan:

prof. dr. Nikolaj Zimic

IZJAVA O AVTORSTVU

diplomskega dela

Spodaj podpisani Matej Žniderič, z vpisno številko **63080178**, sem avtor diplomskega dela z naslovom:

Spletni iskalnik podatkov o osebah

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal/-a samostojno pod mentorstvom doc. dr. Dejana Lavbiča
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., angl.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki »Dela FRI«.

V Ljubljani, dne 20. 9. 2013

Podpis avtorja:

*Ob tej priložnosti se zahvaljujem mentorju doc. dr. Dejanu Lavbiču za prijaznost
in strokovno pomoč pri pisanju diplomske naloge.*

*Zahvaljujem se sestri Tini za nasvete in pomoč pri izdelavi diplomske naloge,
bratu Tomažu in sestri Sari za zabavna druženja med vikendi.*

*Rad bi se zahvalil vsem sošolcem in prijateljem, s katerimi smo se skupaj učili,
delali in zabavali.*

Posebna zahvala gre staršema za podporo in razumevanje med študijem.

Kazalo

Seznam uporabljenih kratic

Povzetek

Abstract

1	Uvod.....	1
2	Spletno luščenje podatkov.....	3
2.1	Tehnike luščenja	3
2.2	Legalnost luščenja.....	4
3	Možnosti luščenja podatkov o osebah v Sloveniji	7
3.1	Telefonski imenik Slovenije	7
3.2	Poslovni imenik bizi.si.....	8
3.3	Finance.si	10
3.4	COBBIS	10
3.5	SICRIS	11
3.6	Wikipedija.....	13
3.7	Statistični urad Republike Slovenije.....	14
3.8	ZZZS	15
3.9	Odvetniška zbornica Slovenije	16
3.10	Dražbe.....	16
3.11	GURS in seznam neplačnikov	17
3.12	Socialna omrežja.....	17
3.13	Portali z novicami	19
3.14	Seznami storitev.....	19
3.15	Spletni oglasniki	19
3.16	Forumi.....	20
3.17	Konceptualni model.....	20
4	Zasnova projekta	23
4.1	Ideja	23

4.2	Možni scenariji integracije podatkov	23
4.3	Obstoječe rešitve	23
4.4	Karakteristike naše aplikacije.....	25
5	Implementacija iskalnika	27
5.1	Uporabljene tehnologije	27
5.2	Izvorna koda aplikacije	28
5.3	Zgradba aplikacije	28
5.4	Uporabniški vmesnik.....	29
5.5	Luščenje iz Telefonskega imenika Slovenije	30
5.6	Luščenje podatkov iz ZZZS	31
5.7	Luščenje podatkov iz časopisa Finance.....	33
5.8	Implementacija funkcije poznanstva	34
5.9	Integracija podatkov	35
5.10	Prilagoditev za šumnike	35
5.11	Prikaz uporabe in primeri poizvedb	37
6.	Sklepne ugotovitve.....	41
	Literatura.....	43

Slike

Slika 1: rezultati iskanja po TIS-u	8
Slika 2: rezultat iskanja po bizi.si	9
Slika 3: rezultat iskanja po strani finance.si	10
Slika 4: rezultat iskanja po COBBIS-u	11
Slika 5: rezultat iskanja po SICRIS-u	12
Slika 6: primer strani na Wikipediji	14
Slika 7: pogostost imena na stran stat.si	15
Slika 8: delnos strukturirani podatki na odvetniški zbornici	16
Slika 9: primer seznama neplačnikov	17
Slika 10: konceptualni model naštetih virov	21
Slika 11: konceptualni model naše aplikacije	26
Slika 12: zgradba aplikacije	29
Slika 13: osnovna stran aplikacije	30
Slika 14: Chrome developer tools	30
Slika 15: Excel datoteka iz ZZZS	32
Slika 16: iskalnik strani finance.si	34
Slika 18: primer iskanja odvetnika	37
Slika 19: primer iskanja za zdravnika	38
Slika 20: primer iskanja raziskovalca	38
Slika 21: skupni članki dveh oseb	39

Seznam uporabljenih kratic

ARPANET (Advanced Research Projects Agency Network) Prvo paketno komutirano omrežje in predhodnik interneta.

HTML (HyperText Markup Language) Označevalni jezik za opis strukture spletnih strani.

XHTML (Extensible HyperText Markup Language) Označevalni jezik za opis strukture spletnih strani, ki ustreza standardu XML.

XML (Extensible Markup Language) Označevalni jezik, berljiv s strani človeka in s strani računalnika.

XPath (XML Path Language) Poizvedovalni jezik za naslavljanje delov dokumenta XML.

API (Application programming interface) Aplikacijski programski vmesnik.

DOM (Document Object Model) Jezik za interakcijo z objekti HTML, XHTML in XML.

HTTP (HyperText Transfer Protocol) Protokol za prenos informacij na spletu.

TRR Transakcijski račun.

COBBIS (Co-operative Online Bibliographic System and Services) Slovenski knjižnični informacijski sistem.

IZUM Inštitut informacijskih znanosti

SICRIS (Slovenian Current Research Information System) Informacijski sistem o raziskovalni dejavnosti v Sloveniji.

ZZZS Zavod za zdravstveno zavarovanje Slovenije.

CSS (Cascading Style Sheets) Jezik za opisovanje videza spletnih strani.

TIS Telefonski imenik Slovenije.

PIRS Poslovni informator Republike Slovenije.

AJPES Agencija Republike Slovenije za javnopravne evidence in storitve.

GURS Geodetska uprava Republike Slovenije.

MOSS (Merjenje obiskanosti spletnih strani) Valutna medijska raziskava za analizo slovenskega spletnega trga.

ERGO (European Research Gateways Online) Evropski informacijski sistem za raziskovalno dejavnost.

Povzetek

Svetovni splet je eno najbolj zanimivih in uporabnih orodij današnjega sveta. Različni spletni viri vsebujejo ogromno količino informacij, ki so koristne na najrazličnejših področjih. Te vire je poleg osnovne uporabe možno luščiti in integrirati. S povezavo različnih virov se lahko pridobijo nove uporabne informacije in izdelajo različne uporabne aplikacije.

Osrednja tema diplomskega dela so različni spletni viri ter možnosti njihovega luščenja in integracije. Osredotoča se na vire, ki ponujajo podatke o osebah v Sloveniji. Narejen je obsežen pregled slovenskih virov, ki vsebuje informacije o viru, podatke, ki jih vir ponuja, in tehniko luščenja, ki je potrebna za pridobitev teh podatkov. Možnih integracij virov je več in so opisani v diplomu. Ena izmed možnih integracij virov je realizirana v obliki spletne aplikacije. Implementiran je iskalnik, ki na podlagi imena in priimka po različnih spletnih virih išče informacije o določeni osebi. Pridobljene podatke med seboj integrira in prikaže na enem mestu. Dodatno na podlagi analize člankov prikaže ljudi, ki se v člankih pojavljajo skupaj z iskano osebo.

Ključne besede

Spletna aplikacija, Java EE, luščenje podatkov, integracija spletnih virov, delno strukturirani podatki

Abstract

The World Wide Web is one of the most interesting and useful tools in today's world. Various online resources contain a huge amount of information that is useful in various areas. In addition to the basic capabilities these resources can be data mined and integrated together. With integration of different sources, it is possible to get new useful information and produce a variety of useful applications.

The central theme of the thesis are the various online resources and possibilities of their mining and integration. It focuses on the resources that offer information on persons in Slovenia. There is a comprehensive overview of Slovenian data sources. Overview includes information about the source, the information provided by the source and technique required for the acquisition of these data. There is more than one possibility of data integration and they are described in the thesis. One of the possible integration of resources is then realized in the form of web application. Search engine that takes first and last name as input and searches across various online sources for information about that person is implemented. Acquired data from different data sources is then integrated and shown in on place. In addition, based on the analysis of articles, application shows people who appear in the articles together with the searched person.

Keywords

Web application, Java EE, web scraping, web resources integration, semi-structured data

1 Uvod

Začetki interneta segajo v šestdeseta leta prejšnjega stoletja, ko se je začel razvoj predhodnika interneta, imenovanega ARPANET. Prava revolucija pa je nastopila s prihodom svetovnega spleta leta 1990, ko so bili izdelani prvi spletni strežnik, spletni brskalnik in spletna stran [1]. V naslednjih dveh desetletjih je internet v celoti zaznamoval svet. V letu 2013 ima tako dostop do interneta in s tem svetovnega spleta dve milijardi sedemsto milijonov ljudi [2].

Danes uporabljamo internet tako rekoč povsod: na delovnih mestih, v učilnicah in v prostem času. Splet nam ponuja veliko koristnih funkcionalnosti, kot so dostop do elektronske pošte, spletno bančništvo in spletno nakupovanje. Poleg tega predstavlja obsežen vir najrazličnejših informacij, kot so podatki o vremenu in novice, ter vir znanj, ki jih potrebujemo med študijem ali v službi.

Skupaj s povečevanjem števila uporabnikov spleta pa se je povečeval tudi njegov obseg. Informacije, ki so bile še desetletje nazaj dostopne le v fizični obliki, so se preselile na splet. Tako so se na splet preselili različni imeniki, časopisi, oglasi, s pomočjo forumov in socialnih omrežij pa tudi ljudje sami. Na spletu objavljamo podatke o sebi, svojem življenju, svoje fotografije in videoposnetke. Narava interneta, ki omogoča, da je vsak uporabnik tudi avtor, pa je prispevala še k dodatnemu povečanju količine spletnih informacij. Debate, ki so se včasih dogajale v zaprti družbi, so se preselile na forume. Z uporabo pametnih mobilnih telefonov pa se je prisotnost na spletu še povečala. Tako ima velik del ljudi praktično ves čas dostop do interneta. Vse spletne strani pa so poleg osnovne uporabe tudi zelo zanimiv vir podatkov. Te vire je možno luščiti in jih uporabiti za izdelavo različnih aplikacij.

Večina virov na spletu je v obliki dokumentov HTML in XML ter predstavlja delno strukturiran vir podatkov. Take dokumente računalnik razume in je iz njih možno programsko pridobiti podatke. V nekaterih primerih pa so lahko dokumenti v računalniku nerazumljivi obliki ali v obliki slik, poleg tega so nekatere strani namenoma izdelane na tak način, da je do njih težko računalniško dostopati. Nas bodo zanimale tehnike luščenja podatkov iz različnih tipov dokumentov. Za učinkovito luščenje pa je treba poznati tudi tehnike blokiranja spletnih robotov, zato bomo pregledali tudi te.

Z vsem tem pa seveda ne želimo kršiti zakonodaje in hočemo ostati v legalnem okviru. Zato je potrebna dobra preučitev svetovne, predvsem pa slovenske zakonodaje. Luščenje podatkov poznamo zelo malo časa in zato tudi zakonodaja še ni čisto definirana. Treba je biti pazljiv in se držati nekih pravil. S pravilno izbiro tehnik luščenja je možno dobiti željen rezultat in ostati znotraj legalnih okvirov. Zanimalo nas bo, kako se tega lotiti in kje je treba biti previden.

Spletnih virov podatkov je ogromno, mi pa se bomo osredotočili na tiste, ki ponujajo podatke o fizičnih osebah v Sloveniji. Uporaba spleta je v Sloveniji zelo razširjena in je tako prisotnih veliko virov, ki ponujajo informacije. Razlogov za pridobivanje podatkov o osebah je več. Lahko gre za preproste stvari, kot je izvedeti, kje stanuje naš prijatelj ali kje je zaposlen. Lahko pa se nam zgodi, da moramo izbrati zdravnika, odvetnika ali drugega izvajalca storitev. Zanimalo nas bo čim več informacij o njih in njihovi usposobljenosti. Prav tako pogosto sodelujemo tudi z ljudmi, s katerimi navežemo stike preko interneta. Zagotovo bi radi izvedeli nekaj o njih in preverili, ali so vredni zaupanja.

Vse dostope do različnih virov lahko nadgradimo z dobro integracijo različnih podatkov. Računalnik omogoča hiter dostop do podatkov in lahko posledično obdela veliko količino podatkov. Z dobro programsko logiko pa je možno te podatke med seboj povezati in s tem pridobiti nove informacije, ki bi bile brez pomoči računalnika težko dostopne. Z integracijo lahko tako na enem mestu dobimo podatke, kot so telefonska številka, naslov, služba, poleg tega pa tudi mnenje medijev, mnenje ljudi in njihovo zadovoljstvo s storitvami. Računalnik nam olajša delo dolgotrajnega iskanja, prihrani čas ter ponudi bolj uporabne informacije.

Z izdelavo spletne aplikacije se bomo lotili prav teh problemov ter luščenje in povezovanje pokazali v praksi. V programskem jeziku Java EE bomo sprogramirali spletni iskalnik. S pomočjo različnih tehnologij bomo dostopali do več virov podatkov, jih med seboj združili in prikazali na enem mestu. Izbrali bomo različne tipe virov, kot so imenik, spletni časopis in spletne zbirke podatkov. Po imenu in priimku bo tako mogoče poiskati uporabne informacije o osebi. Poleg osnovnega iskanja bomo implementirali še dve dodatni funkciji in tako pokazali, kaj je mogoče narediti z različnimi viri. Aplikacija pa bo služila tudi kot osnova za nadaljnji razvoj iskalnikov po osebah.

V nadaljevanju je v drugem poglavju predstavljeno spletno luščenje podatkov in različne tehnike, ki se uporabljajo pri tem. Opisana je slovenska zakonodaja na tem področju in okviri, ki se jih je treba držati, da je ne kršimo. V tretjem poglavju je narejen pregled različnih virov podatkov v Sloveniji in predstavljene možnosti luščenja teh virov. Vsak vir je opisan in predstavljeno je, katere podatke vsebuje in katere tehnike so potrebne za njihovo luščenje. Četrto poglavje je namenjeno predstavitvi različnih scenarijev integracije podatkov. Opisanih je nekaj obstoječih rešitev, nato pa je predstavljen predlog naše rešitve. V petem poglavju je opisana implementacija iskalnika oseb; natančno sta opisana zgradba in delovanje aplikacije. Prikazano je, kako se v praksi dostopa do različnih servisov in kako različne podatke med seboj združiti. Na koncu, v šestem poglavju, so predstavljene zaključne ugotovitve in predlogi možnih izboljšav aplikacije.

2 Spletno luščenje podatkov

Spletno luščenje podatkov je proces pridobivanja delno strukturiranega dokumenta s spleta, običajno je to spletna stran, napisana v jeziku HTML in XHTML, ter analiza tega dokumenta na način, da iz njega pridobimo določene podatke, ki jih potem uporabimo za drug namen [3]. Za razliko od spletnih storitev, katerih namen je računalniško pridobivanje in izmenjava podatkov prek spleta [4], gre pri spletnem luščenju za pridobivanje podatkov iz dokumentov, ki so namenjeni spletnim brskalnikom. Spletni brskalniki so namenjeni prikazovanju dokumentov ljudem. Tako so dokumenti, ki so za ljudi razumljivi, lahko za računalnik popolna neznanka. Poleg tega so brskalniki pri prikazovanju dokumentov, ki se ne držijo konvencij, popustljivi. Za računalnik to postane problem, s tem pa spletno luščenje pogosto postane težavno ali celo nemogoče [3, p. 2].

Nameni luščenja so različni. Vedno več spletnih strani ponuja podatke preko spletnih storitev. V primerih, ko te niso implementirane ali pa nam ponujajo le del podatkov, nam kot edini vir podatkov ostane luščenje. V nekaterih primerih različni sistemi svojih podatkov ne ponujajo preko storitev, ponujajo pa jih svojim uporabnikom preko spleta. Z luščenjem lahko te podatke pridobimo in jih integriramo v zunanji sistem. Tako lahko luščenje postane alternativa spletnim storitvam.

Druga uporaba luščenja je pri izdelavi spletnih pajkov, ki se sprehajajo po spletu in ga indeksirajo. Včasih je izdelava iskalnikov potekala ročno, zdaj pa s pomočjo spletnega luščenja računalniki sami indeksirajo splet in s pomočjo algoritmov uporabnikom ponudijo možnost iskanja po njem.

Spletno luščenje se s simulacijo uporabnika uporablja tudi kot eden izmed načinov avtomatičnega testiranja spletnih aplikacij [3, p. 3].

2.1 Tehnike luščenja

Izreži in prilepi

Strani so ustvarjene za ljudi in včasih tudi najboljši računalniški programi ne morejo izluščiti podatkov. V nekaterih primerih lastniki strani ne želijo, da bi se iz njihove strani luščili podatki, in zato namenoma ustvarijo stran, iz katere podatkov ni mogoče preprosto izluščiti.

XML Path Language ali XPath

XML Path language ali XPath je poizvedovalni jezik za naslavljanje delov dokumenta XML. Poleg tega podpira računske operacije z vrednostmi v dokumentih XML. Rezultat poizvedbe XPath je seznam vozlišč ali vrednost nekega podatka [5].

Regularni izrazi

Regularni izrazi so zaporedje navadnih znakov in nadomestnih znakov, ki imajo poseben namen. Program primerja regularni izraz s tekstom in vrne ujemanja. Zelo nam pridejo prav pri delu in urejanju besedil, omogočajo pa nam tudi luščenje podatkov. S pomočjo pravilno napisanega regularnega izraza se računalnik sprehodi po dokumentu in izloči za nas zanimive podatke [6].

DOM

Document Object Model (DOM) je specifikacija, ki neodvisno od jezika in platforme definira, kako so podatki prikazani in kako z njimi manipulirati [7]. Pri luščenju podatkov iz ciljne spletne strani izdelamo drevo DOM z enim vrhnjim elementom »root«. Ta vsebuje vozlišča (nodes oz. children), ki imajo lahko dodatna vozlišča, listi drevesa pa vsebujejo podatke [8]. Pri luščenju lahko tako s pomočjo DOM-a programsko pridemo do dela spletne strani, ki nas zanima, in izluščimo potrebne podatke.

Orodja za avtomatično testiranje

Orodja za avtomatično testiranje spletnih strani so dober pripomoček razvijalcem strani. Pripomorejo k izboljšanju zanesljivosti rešitve in pomagajo pri iskanju napak. Poleg tega ponujajo objektivni in neodvisen vpogled v delovanje programske rešitve. V spletnem prostoru so to na primer Selenium [9], HtmlUnit [10] in Sahi [11]. Za brskalnik se programi ne razlikujejo od uporabnika. Namesto da ročno pošiljamo vse zahteve HTTP, lahko z njimi simuliramo uporabnika in tako izpolnjujemo obrazce, klikamo na povezave ter si tako olajšamo delo.

Z uporabo računalniškega vida

Pojavljajo se tudi ideje za luščenje podatkov s pomočjo računalniškega vida. Gre za identifikacijo in ekstrakcijo podatkov z vizualno interpretacijo strani, kot jo vidi človek. Tehnika se uporablja predvsem v primerih, ko podatki niso v pravem formatu ali pa želijo lastniki strani preprečiti luščenje.

2.2 Legalnost luščenja

Zakonodaja glede luščenja podatkov tako v tujini kot v Sloveniji ni povsem jasna. Spletne strani pogosto definirajo pogoje uporabe in ne dovolijo luščenja podatkov. V realnosti pa je uveljavljanje teh pravil pred zakonom vprašljivo.

Slovenska zakonodaja je na tem področju slabo definirana. V 221. členu Uradnega lista Republike Slovenije se obravnava napad na informacijski sistem:

- (1) Kdor vdre v informacijski sistem ali kdor neupravičeno prestreže podatek ob nejavnem prenosu v informacijski sistem ali iz njega, se kaznuje z zaporom do enega leta.
- (2) Kdor podatke v informacijskem sistemu neupravičeno uporabi, spremeni, preslika, prenaša, uniči ali v informacijski sistem neupravičeno vnese kakšen podatek, ovira prenos podatkov ali delovanje informacijskega sistema, se kaznuje za zaporom do dveh let.
- (3) Poskus dejanja iz prejšnjega odstavka je kazniv.
- (4) Če je z dejanjem iz drugega odstavka tega člena povzročena velika škoda, se storilec kaznuje z zaporom od treh mesecev do petih let [12].

Tako se po navadi lastniki, ki želijo preprečiti luščenje podatkov, zatečejo k tehničnim rešitvam blokiranja spletnih robotov. Nekatere tehnike so:

- Blokiranje IP-naslova.
- Onemogočitev spletnih API-jev.
- Nekateri roboti se med luščenjem identificirajo in jih je na podlagi tega možno blokirati.
- Blokiranje na podlagi spremljanja prevelike uporabe strani.
- Uporaba sistema preverjanja, ali gre za osebo, kot je CAPTCHA.
- Detekcija s pastmi za robote oz. honeypoti. Ena izmed vrst honeypotov so lahko namenske strani, do katerih robotom dostop ni dovoljen. Roboti, ki pravil ne upoštevajo in kljub temu dostopajo do teh strani, se identificirajo in pustijo svoj IP-naslov. Na podlagi tega se lahko robota nato blokira.
- Uporaba tehnologije CSS sprites, ki skupek slik shrani v eno sliko. Zaupni podatki se shranijo v obliki ene slike in nato s pomočjo tehnologije CSS razrežejo in razporedijo na prava mesta na strani.

Treba je upoštevati, za kakšen namen so bili podatki zbrani in ponujeni na spletu. Pri AJPES-u se zbirajo podatki o poslovnih subjektih, pri čemer je eden izmed podatkov tudi lastnik podjetja. Če bi te podatke integrirali s TIS-om, bi kršili pravila, saj podatki na AJPES-u niso bili zbrani za evidenco fizičnih oseb. Pri povezovanju različnih virov je torej treba biti previden. Lastniki strani ne želijo, da bi se njihovi podatki uporabljali za analizo ljudi in tako razkrivali preveč informacij o osebi. To bi lahko sprožilo nezadovoljstvo z njihovo storitvijo in zahteve za umaknitev določenih podatkov. Poleg tega ne želijo, da bi se njihovi podatki uporabljali za komercialno uporabo. Pri prikazu podatkov je treba jasno prikazati, da podatki niso naši, in navesti vir podatkov.

V našem prototipu se bomo skušali izogniti nezaželenim in nedovoljenim praksam. Naša aplikacija bo samo orodje za pomoč pri iskanju in ne bo ponujala zaupnih informacij.

Uporabljeni viri so javni in dostopni vsem. Z njimi ne bo možno pridobivati zasebnih podatkov osebe. Podatki se ne bodo shranjevali lokalno, ampak se bo med uporabo dostopalo do različnih spletnih virov. Rešitev ne bo dostopna kot javna storitev. Cilj aplikacije je raziskovalne narave, njen namen pa je prikaz različnih načinov luščenja in integracije. Prikazani rezultati bodo imeli tudi povezave do originalnih virov podatkov.

3 Možnosti luščenja podatkov o osebah v Sloveniji

Dostop do interneta ima v svetu vedno več ljudi. Slovenija na tem področju ni izjema. Tako je imelo v prvem četrtletju 2012 dostop do interneta 74 % gospodinjstev, večinoma širokopasovne povezave. Med gospodinjstvi z otroki, v katerih je vsaj en otrok star 15 let, jih je imelo dostop do interneta 95 %. Temu se prilagajo tudi podjetja in drugi ponudniki storitev, ki svoje produkte selijo na splet. Na splet pa se selijo tudi ljudje sami. V letu 2012 je bilo tako na družabnih omrežjih 40 % Slovencev, spletne forume pa jih je bralo kar 50 %. Pri mladih je ta številka še veliko višja [13].

Z večjo uporabo forumov in socialnih omrežij ter s prihodom novih spletnih strani se posledično večja tudi obseg koristnih informacij, do katerih lahko dostopamo. Preko različnih virov lahko dostopamo do aktualnih novic ter dobimo najrazličnejše informacije, kot so podatki o vremenu, cenah, borznih tečajih in ljudeh. Za nas bodo najbolj zanimive informacije o osebah v Sloveniji, ki so na spletu prosto dostopne. Nekateri izmed teh podatkov so najprej obstajali v tiskani obliki in se kasneje preselili tudi na splet, nekateri pa so bili ustvarjeni neposredno v spletnem prostoru.

Glavni viri, iz katerih lahko dobimo uporabne informacije, so podrobneje predstavljeni v nadaljevanju.

3.1 Telefonski imenik Slovenije

Univerzalni telefonski imenik Slovenije (TIS), ki je bil včasih na voljo le v tiskani obliki, je od leta 1996 javno dostopen na naslovu www.itis.si. Od knjižne oblike se razlikuje tudi v tem, da so podatki na spletni strani vedno ažurni, saj jih dnevno posodablajo. Imenik vodi podjetje TSmedia, d. o. o., in ima zanj izključne pravice.

Telefonski imenik vsebuje podatke o telefonskih naročnikih vseh večjih slovenskih operaterjev; vsebuje podatke oseb, ki objavi ne nasprotujejo.

TIS vsebuje več kot 1.000.000 telefonskih števil, je hitro in enostavno dostopen ter nam tako služi kot pomemben vir informacij, kar se kaže tudi v obiskanosti spletne strani. Po podatkih raziskave MOSS (Merjenje obiskanosti spletnih strani) ima TIS deseto najbolj obiskano spletno stran. Raziskava iz leta 2012 kaže, da najmanj nekajkrat letno telefonski imenik uporablja vsak drugi spletni uporabnik v Sloveniji. Telefonski imenik uporablja tedensko ali pogosteje 14 % slovenskih spletnih uporabnikov, mesečnih uporabnikov je 18,9 %, 19,5 % pa je spletnih uporabnikov, ki z namenom uporabe spletnega telefonskega imenika do spleta dostopajo nekajkrat letno [14].

Spletna aplikacija nam omogoča pridobitev naslednjih podatkov o osebi:

- ime in priimek,
- naslov,
- telefonska številka.

Podatki so na voljo v delno strukturirani obliki.

Rezultati iskanja Omeji po: Dejavnosti • Kraju

Kdo, kaj > **Janez novak** x

PODJETJA – 6 **OSEBE – 76**

NOVAK JANEZ	Seunigova ulica 2, Ljubljana 1211 Ljubljana - Šmartno	01 511 64 59	VEČ >
NOVAK JANEZ	Cesta II. grupe odredov 25, Ljubljana 1261 Ljubljana - Dobrunje	01 542 98 36	VEČ >
NOVAK JANEZ	Tugomerjeva ulica 14, 1000 Ljubljana	01 505 31 06	VEČ >

Slika 1: Rezultati iskanja po TIS-u

3.2 Poslovni imenik bizi.si

Poslovni imenik bizi.si v lasti podjetja TSmedia, d. o. o., vsebuje podatke o slovenskih podjetjih; vsebuje kontaktne, finančne in poslovne podatke. Po slovenski zakonodaji so nekateri podatki o podjetjih, kot so prihodki, odhodki in ime lastnika, javni. Agencija republike Slovenije za javnopravne evidence in storitve (AJPES) te podatke objavlja na svoji spletni strani <http://www.ajpes.si>.

Vir podatkov spletne strani bizi.si so predvsem vsi telekomunikacijski operaterji v Sloveniji ter AJPES. Najdemo pa lahko tudi podatke, ki jih urejajo drugi viri, npr. Borza Slovenije, TK-operaterji, Banka Slovenije, Uradni list ali poslovni subjekti sami. Vse te podatke spletna stran zbira in jih urejene ponuja obiskovalcem spletne strani. Predstavlja pomemben vir podatkov za poslovne partnerje, zaposlene osebe ter vsakega, ki bi jih ti podatki utegnili zanimati.

Bizi.si vsebuje podatke 180.000 podjetij. Je najbolj obiskan poslovni spletni imenik v Sloveniji in po podatkih raziskave MOSS (februar 2013) štirinajsta najpogostejša obiskana slovenska spletna stran.

Alternativa s podobnimi karakteristikami je Poslovni informator Republike Slovenije (PIRS), ki je dostopen na www.pirs.si.

Preko iskalnika na strani lahko pridobimo naslednje podatke:

- ime podjetja,
- davčna številka podjetja,
- matična številka,
- status, ali podjetje obratuje,
- ali je podjetje davčni zavezanec,
- datum vpisa v javni register,
- imena zastopnikov,
- TRR podjetja [15].

Podatke lahko pridobimo s pomočjo vgrajenega iskalnika. Imena zastopnikov so prikazana v obliki slike in jih je težko izluščiti. Uporabiti bi bilo treba računalniški vid. Drugi podatki so na voljo v delno strukturirani obliki.

The screenshot shows the Bizi.si search interface. At the top is a navigation bar with links: PODJETJA, E-TRŽNICA, ANALIZATRGA, MOJ BIZI, MOJE PODJETJE, and O NAS. Below this is a green search bar with the text 'Iskanje podjetja:' and a search box containing 'Milan Žniderič'. A red 'Iskanje' button is to the right. Below the search bar are links for 'Napredno iskanje' and 'Iskanje po dejavnosti'. A section titled 'Kriteriji iskanja: Iskanje podjetij (Milan Žniderič) IN Valuta je (EUR)' shows 'Prikazano 1 od 1 rezultatov.' and includes buttons for 'Privzeti pogled (privzeto)', 'Nastavi', 'privzeta razvrstitev', and 'Popravi iskanje'. Below this is a table of search results. The table has columns: 'Izberi: stran / vse Naziv', 'Naslov', 'Pošta', and 'Matična številka'. The first result is for 'POLARIS MILAN ŽNIDERIČ s.p.' with a logo icon, address 'ŠENTJOŠT NAD HORJULOM 71', postal code '1354 HORJUL', and identification number '313379600'.

Izberi: stran / vse Naziv	Naslov	Pošta	Matična številka
 LOGOTIP POLARIS MILAN ŽNIDERIČ s.p.	ŠENTJOŠT NAD HORJULOM 71	1354 HORJUL	313379600

Slika 2: Rezultat iskanja na strani bizi.si

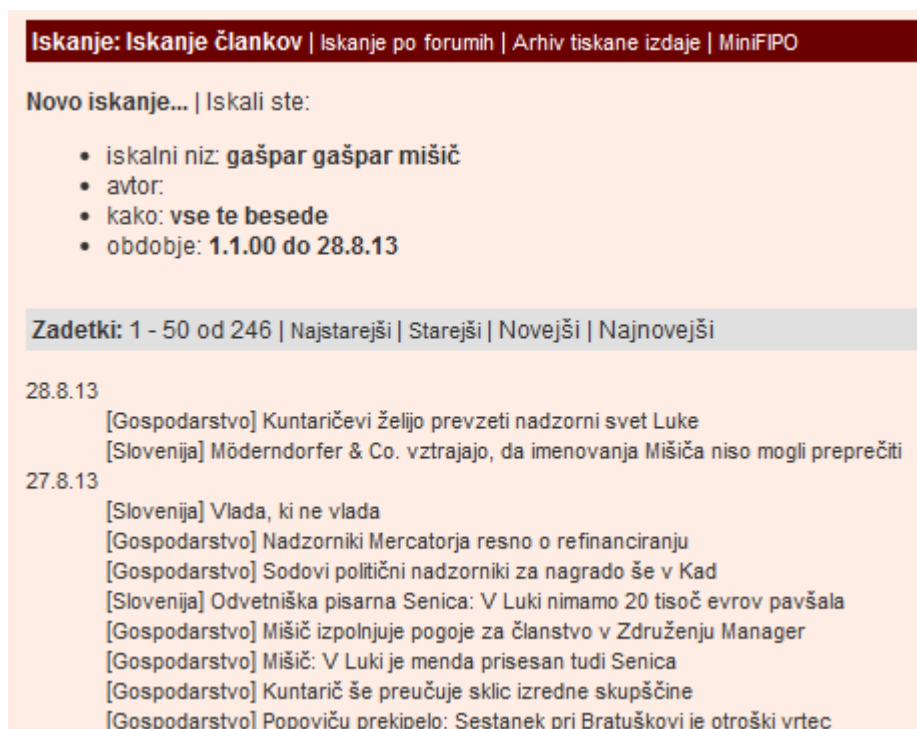
3.3 Finance.si

Časopis Finance je eden izmed najbolj popularnih medijev v slovenskem prostoru, ki se osredotoča na ekonomsko in politično tematiko. Tako kot večina večjih časopisov so se tudi Finance preselile na splet in so dostopne na strani www.finance.si. Spletna stran je bogat vir informacij o ekonomistih, politikih, podjetnikih in drugih znanih osebah v slovenskem prostoru. Na novičarskih portalih veliko ljudi članke tudi komentira in s tem pušča podatke o sebi kot komentatorju, hkrati pa pušča podatke tudi vsebina komentarja [16].

Pridobimo lahko naslednje podatke:

- seznam člankov,
- vsebina člankov,
- seznam ljudi in pravnih subjektov, ki se pojavljajo v člankih,
- datum objave člankov.

Podatki so na voljo preko vgrajenega iskalnika in so v delno strukturirani obliki.



Slika 3: Rezultat iskanja na strani [finance.si](http://www.finance.si)

3.4 COBBIS

COBBIS (Kooperativni online bibliografski sistem in servisi) je slovenski knjižnični informacijski sistem. Je produkt Inštituta informacijskih znanosti v Mariboru (IZUM), njegova funkcija pa je povezovanje podatkov iz knjižnic. Predstavlja platformo nacionalnih

knjižničnih sistemov in bibliografskih sistemov za evalvacijo rezultatov raziskovalnega dela v Sloveniji, Srbiji, Makedoniji, Bosni in Hercegovini, Črni gori, Bolgariji in Albaniji. Skupaj so povezani v regionalno mrežo COBBIS.net. Sistem prispeva k hitrejšemu gospodarskemu in družbenemu razvoju.

Slovenski del COBBIS-a je dostopen na naslovu www.cobiss.si. Ponuja nam iskanje po knjigah, elektronskih medijih, diplomah in drugih gradivih, prisotnih v slovenskih knjižnicah. Prav nam pride pri iskanju in izposoji gradiv. Preko spleta je možno preveriti, ali je gradivo na voljo in kje je na voljo, ter ga tudi rezervirati. Tako nam COBBIS pri delu z gradivi velikokrat olajša delo.

Lahko pa je tudi pomemben vir informacij o osebah. Pridobimo lahko:

- seznam del,
- avtorje del,
- leto izdaje del [17].

Podatki so na voljo preko iskalnika v delno strukturirani obliki.

Rezultati izbirnega iskanja

Iskali ste: Avtor=kajetan kovič IN Izbor zapisov=vse gradivo (tudi e-viri) E

Število najdenih zapisov: 704 SPREMENI ISKALNO

Razvrstitev po: COBISS.SI-ID [zapisi: 1-10] 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 ...

Št.	Avtor	Naslov	Vrsta gradiva	Jezik	Leto	Dostopnost zaloge
1.	Zych, Barbara	Zlata ladja, Koledar 2012 [Slikovno gradivo] : z izbranimi verzi iz otroških pesmi Kajetana Koviča in ilustracijami Jelke Reichman iz slikanice Zlata ladja (MKZ, 2003)	slikovno gradivo	slv	2011	● za izposajo - v čitalnico
2.	Zych, Barbara	Zlata ladja, Koledar 2012 [Slikovno gradivo] : z izbranimi verzi iz otroških pesmi Kajetana Koviča in ilustracijami Jelke Reichman iz slikanice Zlata ladja (MKZ, 2003)	slikovno gradivo	slv	2011	● za izposajo - v čitalnico
3.	Kovič, Kajetan Zlobec, Ciril Menart, Janez	Pesmi štirih	poezija	slv	2003	● za izposajo - na dom

Slika 4: Rezultat iskanja po COBBIS-u

3.5 SICRIS

Informacijski sistem o raziskovalni dejavnosti v Sloveniji (SICRIS; Slovenian Current Research Information System) razvijata in vzdržujeta IZUM in Agencija za raziskovalno dejavnost republike Slovenije (ARRS). Dostopen je na naslovu <http://www.sicris.si>.

Trenutno so v SICRIS-u predstavljene naslednje entitete:

- 958 raziskovalnih organizacij (izpis vseh organizacij),
- 1.508 raziskovalnih skupin (izpis vseh skupin),
- 14.267 raziskovalcev (izpis vseh raziskovalcev),
- 5.713 raziskovalnih projektov (izpis vseh projektov),
- 451 raziskovalnih programov (izpis vseh programov).

Podatki o projektih so integrirani v evropski informacijski sistem za raziskovalno dejavnost (ERGO; European Research Gateways Online), ki povezuje vse evropske informacijske sisteme. V sistemu ERGO je možno iskati po podatkih o raziskovalnih projektih vseh držav hkrati. Na področju pridobivanja podatkov o osebah je najbolj zanimiva obsežna baza raziskovalcev.

Javno ponuja veliko zanimivih podatkov:

- osnovni identifikacijski podatki (ime, priimek, izobrazba, raziskovalna organizacija, skupina),
- podatki za možnost komuniciranja (telefon, telefaks, elektronski naslov, URL),
- podatki o zaposlitvah raziskovalca (organizacija, raziskovalna skupina, datum zaposlitve, delovno mesto, naziv),
- seznam raziskovalnih projektov, v katere je vključen raziskovalec,
- raziskovalno področje in opis ekspertnih znanj (ključne besede, klasifikacija, izobrazba, znanje jezikov),
- reprezentativne bibliografske enote [18].

Podatki so na voljo preko iskalnika v delno strukturirani obliki.

25526	Dr. Lavbič Dejan
STATUS	raziskovalec - aktiven v raziskovalni organizaciji
PREDSTAVITEV	ZAPOSLOTVE
PROJEKTI	PROGRAMI

KONTAKTNI PODATKI

TELEFON	
FAKS	
ELEKTRONSKA POŠTA	dejan.lavbic@fri.uni-lj.si
WWW NASLOV	

DEJAVNOST

↳ **ARRS KLASIFIKACIJA**

2.07.05 - Tehniške vede / Računalništvo in informatika / Informacijski sistemi - programska oprema

↳ **BIBLIOGRAFIJA**

Reprezentativne bibliografske enote / [Osebna](#)

[Citiranost bibliografskih zapisov v COBIB.SI, ki so povezani z zapisi v WoS in/ali Scopus](#)
povezani zapisi: 3/13, citati: 8/18, čisti citati: 7/14, povprečje čistih citatov: 2,33/1,08

Slika 5: Rezultat iskanja po SICRIS-u

3.6 Wikipedija

Wikipedija je brezplačna spletna enciklopedija, dostopna na naslovu www.wikipedia.org. Je eno največjih spletišč, ima 30 milijonov člankov v kar 287 jezikih. Nastaja s sodelovanjem prostovoljcev s celega sveta, ki pišejo članke oz. dodajajo in urejajo vsebino. Brezplačen dostop in velika obsežnost omogočata uporabniku enostavno pridobivanje velikih količin podatkov. Wikipedijo lahko ureja vsak, kar pripomore k njeni bogati vsebini, hkrati pa se moramo zavedati, da moramo biti pri njeni uporabi previdni. Obsežen je tudi slovenski del Wikipedije, ki je dostopen na naslovu http://sl.wikipedia.org/wiki/Glavna_stran in vsebuje več kot 130.000 člankov [19].

Na Wikipediji lahko med drugimi najdemo tudi članke o določenih osebah; pričakujemo lahko, da bomo našli članke in vsebine, povezane predvsem z bolj znanimi osebami. Na Wikipediji lahko običajno pridobimo naslednje podatke o osebah:

- ime in priimek,
- datum rojstva,
- kraj rojstva,
- izobrazba,
- podatki o družini,
- podatki o poklicu,
- poleg tega so običajno prisotni tudi podatki o pomembnih dosežkih, mejnikih v življenju.

Podatki so na voljo preko iskalnika v delno strukturirani obliki.

Ivana Kobilca, slovenska slikarka, * 20. december 1861, Ljubljana, † 4. december 1926.

Vsebina [skrij]

- 1 Življenje
- 2 Dela in pomen
- 3 Literatura
- 4 Zunanje povezave

Življenje [uredi]

Ivana Kobilca se je **risanja** učila pri **Idi Künl** v **Ljubljani**, v letih **1879-1880** pa je v **galeriji dunajske akademije** kopirala slike starih mojstrov. Od leta **1881-1891** je svoje znanje izpopolnjevala v **Münchnu**, naslednjih 7 let pa še pri **žanristu** in **portretistu Aloisu Erdtelu**. Med počitnicami je zbirala in risala **kmečke starožitnosti**. Leta 1891 je na pobudo vzornika **Fritza von Uhdeja** poslala svoji sliki *Poletje* in *Likarice* v **pariški Salon**, kjer ji je Državno umetnostno združenje podelilo naziv *associée*.

V Parizu je živela v letih 1891 in 1892, kjer se je krajši čas izpopolnjevala v zasebni šoli **Henrija Gervexa**, vmes pa je 1892. leta slikala v **Barbizonu**. Leta 1894 je obiskala **Firence**, nato pa zaradi ugodnih naročil v letih 1897-1905 živela v **Sarajevu**, kjer je bila tudi članica Sarajevskega slikarskega kluba. Tu je

Ivana Kobilca



Avtoportret v belem

Rojstno ime:	Ivana Kobilca
Rojstvo:	20. december 1861 Ljubljana, Slovenija
Smrt:	4. december 1926 (64 let) Ljubljana, Slovenija
Narodnost:	 Slovenka
Področje:	slikarstvo, risanje ,

Slika 6: Primer strani na Wikipediji

3.7 Statistični urad Republike Slovenije

Statistični urad Republike Slovenije je glavni izvajalec in povezovalac dela na področju državne statistike. Poleg povezovanja in usklajevanja statističnega sistema sodijo med njegove najpomembnejše naloge še mednarodno sodelovanje, določanje metodoloških in klasifikacijskih standardov, predvidevanje potreb uporabnikov, zbiranje, obdelava in izkazovanje podatkov ter skrb za njihovo zaupnost [20].

Na svoji spletni strani <http://www.stat.si> objavljajo najrazličnejše statistične podatke, kot so stopnja brezposelnosti, povprečna mesečna plača in cene izdelkov, ki lahko pridejo prav podjetjem in državljanom. Na področju ljudi spremljajo podatke o pogostosti imen in priimkov, kar nam lahko pride prav za preverjanje, koliko oseb je v Sloveniji z istim imenom in priimkom. Zaradi zagotavljanja zasebnosti pa v primeru, da je oseb z istim imenom in priimkom manj kot 5, ne ponudijo točne številke.

Podatki so na voljo v obliki člankov v delno strukturirani obliki. Pogostost imen pa je na voljo preko vgrajenega iskalnika.

The screenshot shows the navigation bar of the Slovenian Statistical Office (Statistični urad RS) with links: DRŽAVNA STATISTIKA, STATISTIČNI URAD, PUBLIKACIJE, PRVI OBISK, and KONTAKT. Below the navigation bar, the breadcrumb trail reads: Prebivalstvo > Interaktivna orodja > Baza rojstnih imen in priimkov. The main content area displays the search results for the combination of name and surname 'JANEZ NOVAK', stating it appears 167 times among males. It provides a link for a more detailed view and lists the counts for each part: JANEZ (23,690) and NOVAK (11,258). A section titled 'IZBRANI SPOL' (Selected Gender) explains that the search was performed for males and provides a link to search by gender.

DRŽAVNA STATISTIKA STATISTIČNI URAD PUBLIKACIJE PRVI OBISK KONTAKT

Prebivalstvo > Interaktivna orodja > Baza rojstnih imen in priimkov

Kombinacija imena in priimka **JANEZ NOVAK** se med moškimi pojavlja **167-krat**.

Za natančnejši pregled kliknite na izbrano ime ali priimek:

- **JANEZ** - število moških s tem imenom: **23.690**
- **NOVAK** - število oseb s tem priimkom: **11.258**

IZBRANI SPOL

Izbrano kombinacijo imena in priimka ste iskali med **moškimi**. Če bi ga želeli iskati med ženskami, kliknite na spodnjo povezavo.

» Sprememba spola

Slika 7: Pogostost imena na strani Statističnega urada RS

3.8 ZZZS

Zavod za zdravstveno zavarovanje (ZZZS) je javni zavod za izvajanje obveznega zdravstvenega zavarovanja. Funkcija zavoda je zagotavljanje učinkovitega zbiranja in razdeljevanja javnih sredstev. Pravice iz obveznega zdravstvenega zavarovanja, za katere se namenja sredstva, zbrana z obveznim plačevanjem prispevkov, zajemajo pravice do zdravstvenih storitev in nekaterih denarnih nadomestil. Na spletni strani www.zzzs.si je možno najti veliko informacij s področja zdravstvenega zavarovanja, pravic, prispevkov. Izvajalcem zdravstvenih storitev portal omogoča še dostop do cenikov, baze zdravil in elektronskega poslovanja. Javno objavlja tudi čakalne dobe v posameznih ambulantah, cenike in podatke o zdravnikih [21].

Za nas so najbolj zanimivi prav podatki o zdravnikih. Na uradni spletni strani so dostopne tri datoteke, in sicer podatki o osebnih zdravnikih, zobozdravnikih in ginekologih.

V datotekah so na voljo naslednji podatki:

- ime in priimek,
- vrsta dejavnosti oz. zdravstvene storitve,
- lokacija dela.

Na voljo so v Excelovi datoteki in so v strukturirani obliki.

3.9 Odvetniška zbornica Slovenije

Odvetniška zbornica Slovenije je organizacija odvetnikov, ki spremlja in obravnava problematiko dela odvetnikov ter skrbi za razvoj odvetništva. Na svoji spletni strani na naslovu <http://www.odv-zb.si/imenik/imenik-odvetnikov> objavlja različne informacije, koristne za odvetnike. Na strani je tudi iskalnik vseh odvetnikov v Sloveniji.

S pomočjo iskalnika lahko pridobimo naslednje podatke o odvetnikih:

- ime in priimek,
- naslov odvetniške pisarne,
- kraj odvetniške pisarne,
- službena številka.

Podatki so na voljo preko vgrajenega iskalnika v delno strukturirani obliki [22].

Rezultati iskanja

Izvozi

Klikni na odvetnika za podroben ogled

Ime	Priimek	Naslov	Kraj	Telefon
Janez	Ahlin	Kolodvorska ulica 7	Ljubljana	01 232 26 22
[1]				

na vrh

Slika 8: Delno strukturirani podatki na strani Odvetniške zbornice Slovenije

3.10 Dražbe

V Sloveniji javne dražbe ponujajo:

- carina na naslovu http://www.carina.gov.si/si/javne_objave/javne_drazbe,
- policija na naslovu <http://www.policija.si>,
- ministrstvo za pravosodje na naslovu http://www.mp.gov.si/si/o_ministrstvu/javne_objave/javne_drazbe.

V spletnih objavah ni veliko informacij o lastniku dražbenega artikla ali o udeležencih dražbe, vendar pa bi se lahko ti podatki v kombinaciji z drugimi viri podatkov izkazali za uporabne.

Pridobimo lahko:

- informacije o dražbenem izdelku,
- izklicno ceno,
- v primeru nepremičnine tudi naslov nepremičnine,
- ime organizatorja dražbe.

Podatki so na voljo preko spletnih objav v delno strukturirani obliki.

3.11 GURS in seznam neplačnikov

Geodetska uprava Republike Slovenije (GURS) na svoji spletni strani enkrat mesečno objavi štiri sezname neplačnikov davka: seznam pravnih oseb, fizičnih oseb, davčnih oseb v insolventnih postopkih in davčnih zavezancev, ki niso v insolventnih postopkih. Vsi podatki nam lahko pridejo prav pri poslovanju in preverjanju ozadja podjetja oz. pravne osebe.

Seznam neplačnikov je na strani http://seznam.gov.si/DURS/main_neplacniki.html.

Pridobimo lahko:

- ime podjetja, ki dolguje davke,
- naslov podjetja, ki dolguje davke.

Podatki so prikazani kot slika in jih je zato računalniško težko izluščiti [23].

Neplačniki pravne osebe z začetnico D na dan 25.7.2013

[Prejšnja stran](#) [Vstopna stran](#) [Naslednja stran](#) [Razredi neplačnikov](#)

P02 80809049 D & DD NOVO MESTO, D.O.O.; VRHOVČEVA ULICA 18, NOVO MESTO
P02 11195371 D&D GRADBENIŠTVO D.O.O.; CESTA ANDREJA BITENCA 68, LJUBLJANA
P01 20403038 D.EGO.2010, D.O.O.; ŽIDOVSKA ULICA 12, MARIBOR
P01 46064486 D.H.B., D.O.O.; KOČEVSKA CESTA 119, KOČEVJE
P02 96137541 D.M.N., D.O.O.; CESTA NA BRŠLJANOVEC 12, SEŽANA
P06 18125239 DA & JU, D.O.O.; OB LJUBLJANICI 24, LJUBLJANA
P02 49079140 DABA NOVO MESTO, D.O.O.; STOPIČE 20 A, STOPIČE
P01 68172222 DABIČ D.O.O.; SLOMŠKOVA ULICA 27, AJDOVŠČINA
P03 20250193 DACI D.O.O.; ULICA GRADNIKOVE BRIGADE 53, NOVA GORICA
P02 59075007 DADA, VALENČIČ IN DRUGI TRGOVINA NA DEBELO IN DROBNO, D.N.O.; NAVRATILOVA POT 2, METLIKA
P01 29721695 DADJA D.O.O.; CESTA V REČICO 14, LAŠKO

Slika 9: Primer seznama neplačnikov

3.12 Socialna omrežja

Frendi in flirt ter Ona-on

Večji slovenski socialni omrežji Friendi in Flirt, dostopno na <http://frendiinflirt.24ur.com>, in Ona-on, dostopno na <http://www.ona-on.com>, sta namenjeni spoznavanju ljudi. Ljudje se na strani registrirajo in napišejo nekaj o sebi, vnesejo tudi datum rojstva. Ti podatki so za luščenje lahko zelo zanimivi, vendar se večina ljudi registrira z vzdevkom, zato je težje ugotoviti, za katero osebo v resnici gre.

Facebook

Na največjem svetovnem socialnem omrežju Facebook je bilo februarja 2013 750.000 Slovencev [24]. Večina se registrira s pravim imenom in priimkom. Ljudje na njem delijo slike, sporočila in ogromno podatkov. Za javnost je veliko podatkov skritih, s prijatelji pa delijo datum rojstva, slike, prijateljstva, hobije in ogromno drugih podatkov. Tako je lahko to socialno omrežje zelo pomemben vir informacij o osebah.

Katere podatke lahko pridobimo, je odvisno od nastavitve iskane osebe. V večini primerov pa lahko, če z osebo nismo prijatelji, pridobimo:

- ime in priimek,
- profilno sliko.

Če smo prijatelji z določeno osebo, pa lahko poleg zgornjih podatkov pridobimo veliko dodatnih podatkov. Nekateri izmed njih so:

- datum rojstva,
- seznam prijateljev,
- interesi osebe,
- objave osebe.

Facebook ponuja podatke preko uradnega vmesnika API, ki omogoča hitro in preprosto pridobivanje želenih podatkov. Zaradi zagotavljanja zasebnosti pa vmesnik API ne ponuja vseh vidnih podatkov, poleg tega mora iskana oseba namestiti aplikacijo. Tako se je v nekaterih primerih smiselno lotiti luščenja strani in uporabiti delno strukturirane podatke.

Twitter in LinkedIn

Omrežji Twitter, dostopno na <http://twitter.com>, in LinkedIn, dostopno na <http://www.linkedin.com>, sta podobni Facebooku. Twitter se razlikuje v tem, da so objave na omrežju lahko dolge največ 140 znakov, LinkedIn pa je socialno omrežje, namenjeno sklepanju poslovnih kontaktov. Ljudje na omrežju objavljajo predvsem podatke o izobrazbi, pridobljenih znanjih in delovne izkušnje.

Podatke je v omejeni obliki možno pridobiti preko uradnih API-jev.

3.13 Portali z novicami

Poleg že omenjenega portala finance.si so na voljo tudi drugi novičarski portali. V slovenskem prostoru so to na primer 24ur.com, rtvslo.si in dnevnik.si. Portali so obsežen vir informacij o ljudeh, zanimivi pa so lahko tudi komentarji člankov.

Možno je pridobiti:

- sezname člankov, v katerih je omenjena neka oseba,
- podatke o tej osebi,
- na podlagi komentarjev mnenje javnosti o določeni osebi.

Podatki so na voljo v delno strukturirani obliki.

3.14 Seznami storitev

Na spletu je prisotnih veliko storitvenih portalov, ki zbirajo in kategorizirajo podatke o ponudnikih storitev. V slovenskem prostoru so to na primer <http://www.storitev.si>, <http://rumenestrani.si> in storitveni del spletnega oglasnika Bolha, dostopen na <http://www.bolha.com/storitve>. Namen teh portalov je iskanje in primerjava različnih ponudnikov storitev. Poleg tega so tudi obsežen in zanimiv vir podatkov.

Izvemo lahko:

- vrsto dejavnosti,
- ime izvajalca dejavnosti,
- lokacijo izvajanja dejavnosti.

Podatki so na voljo preko vgrajenih iskalnikov.

3.15 Spletni oglasniki

Večji spletni oglasniki v slovenskem prostoru so splošni oglasnik <http://bolha.com>, oglasnik za prodajo cestnih vozil <http://avto.net>, in oglasnik za prodajo nepremičnin <http://www.nepremicnine.net>.

Ponujajo nam predvsem informacije o prodajnem artiklu in telefonsko številko prodajalca. V kombinaciji z drugimi viri podatkov in s pomočjo spletnih pajkov pa bi bilo mogoče izvedeti tudi nekaj informacij o prodajalcu, številu njegovih oglasov ter zgodovini prodaj.

Pridobimo lahko:

- telefonsko številko prodajalca,
- opis artikla.

Podatki so na voljo v delno strukturirani obliki.

3.16 Forumi

Zaradi obsežnosti podatkov so lahko forumi zelo zanimiv vir podatkov. Bolj znani forumi v slovenskem prostoru so na primer <http://slo-tech.com>, <http://med.over.net> in <http://mobisux.com>.

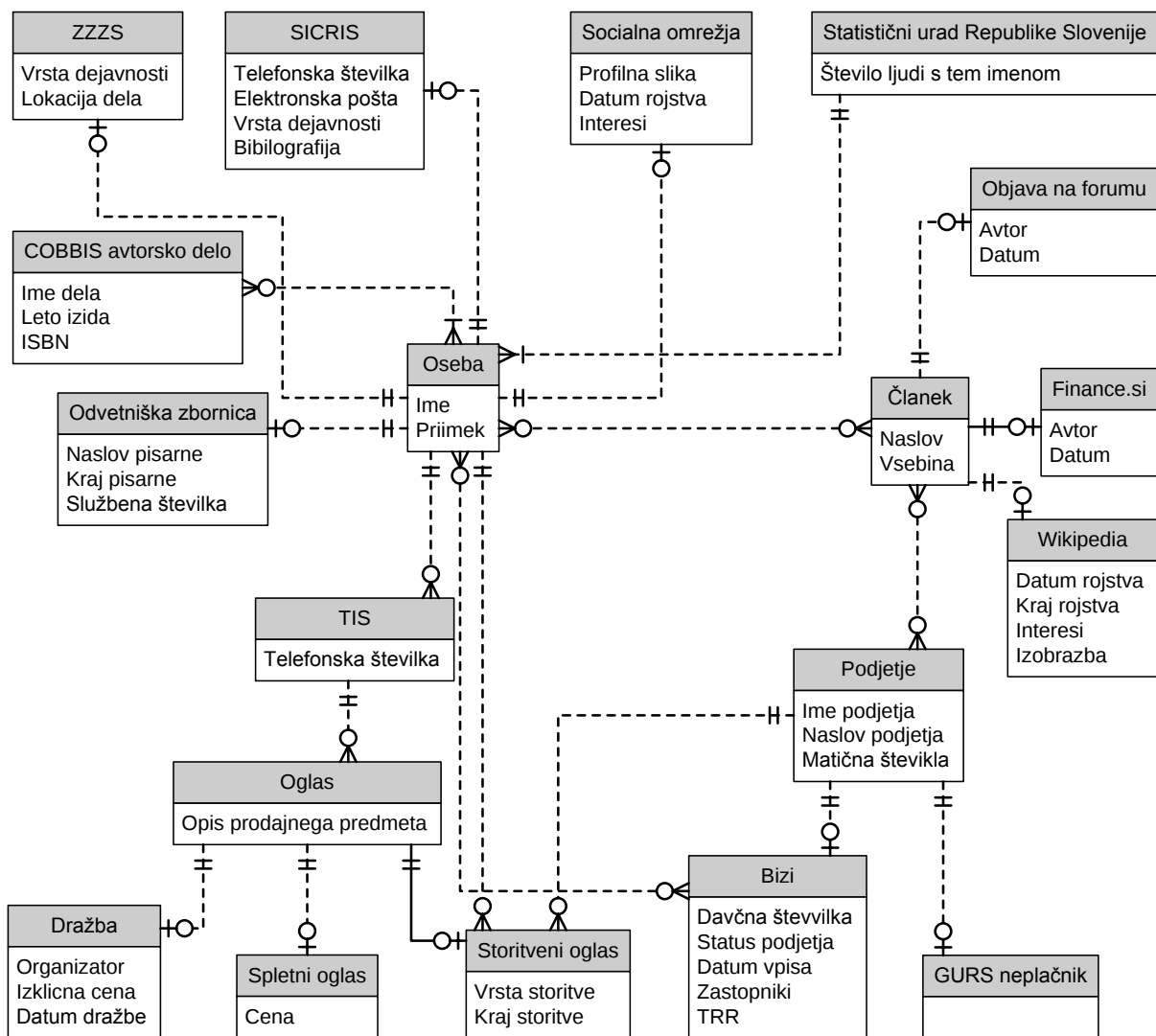
Pridobimo lahko:

- mnenje javnosti o določeni osebi, stvari,
- objavljena sporočila določene osebe.

Problem forumov je, da so podatki na voljo v nestrukturirani obliki v obliki naravnega jezika in jih je računalniško težko izluščiti.

3.17 Konceptualni model

Konceptualni model prikazuje, kako so različne entitete povezane med seboj. Veliko entitet je možno povezati med seboj na podlagi imena in priimka. Problem je, da to ni enolični identifikator in gre lahko za drugo osebo z enakim imenom. V teh primerih bi bilo dobro, da bi lahko entitete povezali še po kakšnem drugem atributu, vendar pa v večini primerov to ni možno.



Slika 10: Konceptualni model naštetih virov

4 Zasnova projekta

4.1 Ideja

V prejšnjem poglavju smo videli, da je v Sloveniji veliko različnih spletnih strani, ki so lahko zelo koristne; nekatere izmed njih lahko poleg osnovne uporabe uporabimo tudi kot vir podatkov. S pravo kombinacijo virov in pametno integracijo podatkov bi se tako dalo izdelati novo aplikacijo, ki bi nam ponujala nove načine uporabe.

4.2 Možni scenariji integracije podatkov

Spletni oglasniki s Telefonskim imenikom Slovenije

Spletni oglasniki vsebujejo informacije o izdelku, ki ga oseba prodaja, in informacije o osebi. Po navadi so to ime, priimek, telefonska številka in kraj prodajalca.

O prodajalcu želimo izvedeti več. V povezavi s TIS-om lahko iz imena osebe pridobimo naslov in telefonsko številko ter obratno, iz telefonske številke lahko ugotovimo, za koga gre. V povezavi s poslovnim imenikom bizi.si lahko ugotovimo, ali ima oseba podjetje. S pomočjo spletnih pajkov pa dobimo tudi število in vrsto izdelkov, ki jih oseba prodaja. Tako lahko ugotovimo, ali gre za vsakdanjega prodajalca ali pa morda za preprodajalca.

Zbirke podatkov o osebah z novicami

Zbirke zdravnikov, odvetnikov, raziskovalcev lahko povežemo z novicami in objavami po forumih. Če je oseba velikokrat omenjena skupaj z drugo osebo, je večja verjetnost, da jo pozna ali z njo sodeluje. Če je oseba velikokrat omenjena skupaj z odvetniki, lahko to pomeni, da je v sodnih preiskavah. Če je oseba velikokrat omenjena skupaj z raziskovalci, pa lahko to pomeni nekaj dobrega.

Kombinacija različnih portalov z novicami, forumi in blogi za oceno mnenja javnosti

Glede na članke in komentarje novic lahko ugotovimo, kaj si ljudje mislijo o:

- določeni osebi,
- določeni skupini ljudi (zdravniki itd.).

S pomočjo teh podatkov lahko predvidevamo, kdo bo zmagal na volitvah. Če ugotovimo, da je v Sloveniji zelo slabo mnenje o zdravnikih, je morda treba narediti nekaj v smeri, da se to izboljša.

4.3 Obstoječe rešitve

Na spletu obstaja več servisov, ki nam omogočajo iskanje po osebah in pridobivanje podatkov iz različnih virov. Delujejo po principu, da v iskalnik napišemo ime in priimek osebe, potem

pa iskalnik iz različnih virov, kot so socialna omrežja, blogi in imeniki, najde ujemanja in jih predstavi na enem mestu.

Popularne aplikacije so na primer: 123people, Pipl, ZabaSearch, Supervizor, Kdo kliče?

123people

123people je eden izmed najbolj obiskanih iskalnikov oseb na svetu. Dostopen je na naslovu <http://www.123people.com> in išče samo po javnih virih informacij. Ponuja svetovno in v primeru večjih držav tudi prilagojeno iskanje [25].

Išče po več kot 200 virih, nekateri izmed njih so:

- socialna omrežja Facebook, Google+, LinkedIn, Twitter,
- blogi,
- telefonski imeniki,
- policijske kartoteke,
- zbirke elektronskih naslovov [26].

Zaradi velike količine virov je lahko zelo koristno orodje. Vendar ne podpira prilagojenega iskanja za Slovenijo in tako ob iskanju oseb iz Slovenije vrne pomanjkljive informacije.

Pipl

Iskalnik Pipl je dostopen na <https://pipl.com> ter poleg iskanja po imenu in priimku ponuja še dodatne funkcije, kot je iskanje po telefonski številki, elektronskem naslovu in uporabniških imenih. Poleg javnih virov išče tudi po »skritem delu« interneta, t. i. deep web oz. invisible web. Deep web je tisti del interneta, ki ga splošnonamenski iskalniki ne morejo doseči. To so na primer:

- dinamične strani, ki so dostopne samo preko določenih poizvedb,
- strani, na katere ne kaže nobena povezava,
- zasebne strani, za katere je potrebno uporabniško ime in geslo,
- strani, do katerih se dostopa s izvajanjem Javascripta in AJAX-a,
- netekstovne datoteke, na primer slike in videi, ki jih standardni spletni pajki ne morejo indeksirati.

Pipl ima prilagojene pajke za brskanje po nekaterih izmed takih strani ter tako vrne več informacij in boljše informacije o ljudeh. Na žalost pa tudi Pipl ni posebej prilagojen slovenskemu okolju in tako ne vključuje vseh slovenskih zanimivih virov [27].

ZabaSearch

Iskalnik ljudi ZabaSearch je dostopen na <http://www.zabasearch.com> ter omogoča iskanje po imenu in priimku, telefonski številki ali po sporočilih, v katerih je oseba omenjena. ZabaSearch vse podatke pridobiva iz javno dostopnih virov [28].

Supervizor

Spletna storitev Komisije za preprečevanje korupcije Supervizor, dostopna na <http://supervizor.kpk-rs.si>, omogoča vpogled v izdatke javnih institucij, ki se nanašajo na blago in storitve. S tem povečuje odgovornost nosilcev javnih funkcij za učinkovito porabo javnih sredstev ter zmanjšuje tveganje za slabo upravljanje in zlorabo oblasti. Podatke o pravnih subjektih pridobiva iz AJ PES-a in iz seznama davčnih zavezancev, ki ga objavi DURS. Podatki o finančnih transakcijah in podatki iz registra proračunskih uporabnikov so pridobljeni s strani Uprave za javna plačila, podatki o javnih naročilih pa s strani Ministrstva za finance in Uradnega lista Republike Slovenije [29].

Aplikacija za mobilne telefone Kdo kliče?

V Sloveniji ni spletne aplikacije za iskanje ljudi, obstaja pa aplikacija Kdo kliče? za mobilne telefone z operacijskim sistemom Android. V primerih, ko nas kliče neznana številka, aplikacija poišče podatke iz PIRS-a in notranje zbirke telefonskih števil ter tako identificira klicatelja in prikaže njegove podatke. Za razliko od naše aplikacije uporablja za iskalni ključ telefonsko številko ter uporablja samo en slovenski vir podatkov (PIRS) [30].

4.4 Karakteristike naše aplikacije

Tudi mi se bomo lotili izdelave iskalnika oseb. Za razliko od prej omenjenih rešitev smo se v prototipu osredotočili na slovenski prostor. Poleg tega bomo poleg osnovnega iskanja implementirali tudi dve dodatni funkciji.

Uporabili bomo naslednje vire:

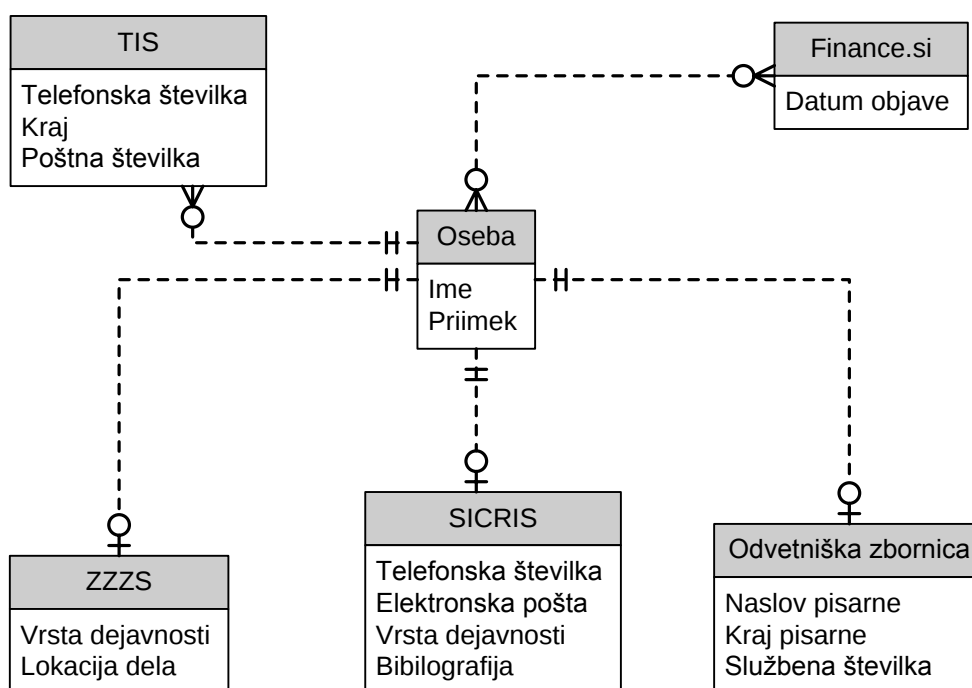
- Telefonski imenik Slovenije,
- Finance,
- Odvetniška zbornica Slovenije,
- ZZSZ,
- SICRIS.

Dostopali bomo do TIS-a ter pridobili nekaj osnovnih podatkov, kot so ime, priimek in telefonska številka.

S pomočjo SICRIS-a, ZZZS-ja in Odvetniške zbornice Slovenije bomo ugotovili, ali je oseba raziskovalec, zdravnik ali odvetnik. Če bo oseba pripadala eni od teh treh skupin, bomo prikazali tudi podatke, ki jih dobimo s teh naslovov.

S pomočjo Financ bomo poiskali članke, v katerih je oseba omenjena. Poleg tega bomo poiskali tudi osebe in pravne subjekte, ki se največkrat pojavljajo skupaj z iskano osebo, ter omogočili prikaz skupnih člankov iskane osebe in dodatne osebe.

Konceptualni model prikazuje, kako bomo podatke povezali med seboj.



Slika 11: Konceptualni model naše aplikacije

5 Implementacija iskalnika

5.1 Uporabljene tehnologije

Razvojno okolje Eclipse

Odprikodno razvojno okolje Eclipse je eno izmed vodilnih okolij. Napisano je večinoma v Javi, uporablja pa se za razvoj aplikacij v Javi in s pomočjo vtičnikov (pluginov) tudi v drugih jezikih, kot so C, C++, PHP, Python, Ruby on Rails. Uporabili smo Eclipse za Javo EE (Eclipse for Java EE developers), ki nam omogoča razvoj spletnih aplikacij [31].

JavaEE

Java Platform, Enterprise Edition ali JavaEE je platforma, ki zagotavlja vmesnik API in razvijalsko okolje za razvoj in izvajanje poslovnih aplikacij. Za prikaz strani smo uporabili tehnologijo Java Server Pages (JSP), za procesiranje zahtev HTTP pa Servlet API.

JBoss AS 7.1

Aplikacijski strežnik, ki ga razvija podjetje Red Hat, je napisan v Javi in tako deluje na vseh operacijskih sistemih, ki podpirajo Javo. Je platforma programskega jezika Java EE. Naloga aplikacijskega strežnika je izdelava dinamičnih spletnih strani. Za JBoss smo se odločili zaradi brezplačnosti in odprtokodnosti, poleg tega pa nam ponuja vse funkcionalnosti, ki jih potrebujemo. Zaradi popularnosti je tudi v primeru težav lahko najti rešitev na internetu [32].

XPath

XML Path language ali XPath je jezik, namenjen dostopanju do določenih delov dokumenta XML. XPath pretvori dokument v drevo vozlišč, nato pa nam omogoči preprost in hiter dostop do delov dokumenta [33].

Knjižnica HtmlUnit

HtmlUnit je spletni brskalnik za programe Java, ki nima grafičnega uporabniškega vmesnika. Omogoča programski dostop do spletnih strani na podoben način, kot to počne uporabnik v navadnem spletnem brskalniku; omogoča izpolnjevanje obrazcev in klikanje povezav. Najbolj pogosta uporaba programa je za avtomatično testiranje spletnih strani, uporabi pa se lahko tudi za spletno luščenje. Za to smo ga uporabili tudi mi [10].

HyperText Markup Language (HTML)

HTML je označevalni jezik, namenjen izdelavi spletnih strani. Opisuje strukturo in vsebino dokumenta. Sestavljen je iz različnih značk, ki predstavljajo različen tip elementov [34].

Kaskadne stilske podloge (CSS)

CSS je jezik za opisovanje videza spletnih strani. Z njim lahko določimo velikost in obliko pisave, postavitev elementov, odmike, barve, omogoča pa tudi nadzor nad aktivnostmi uporabnika. Velika prednost CSS-ja je ločitev oblike od strukture strani, kar omogoča lažje vzdrževanje strani, deljenje stilov z drugimi stranmi in prilagoditev strani različnim okoljem [35].

Chrome developer tools

Chrome developers tools je skupek orodij, ki pridejo skupaj z brskalnikom Chrome ter razvijalcem omogočajo testiranje in razhroščevanje kode. Omogočajo podroben vpogled v delovanje brskalnika in spletne aplikacije. V naši aplikaciji smo jih uporabili za pregled zgradbe delovanja strani, iz katerih želimo izluščiti podatke [36].

Knjižnica Apache POI

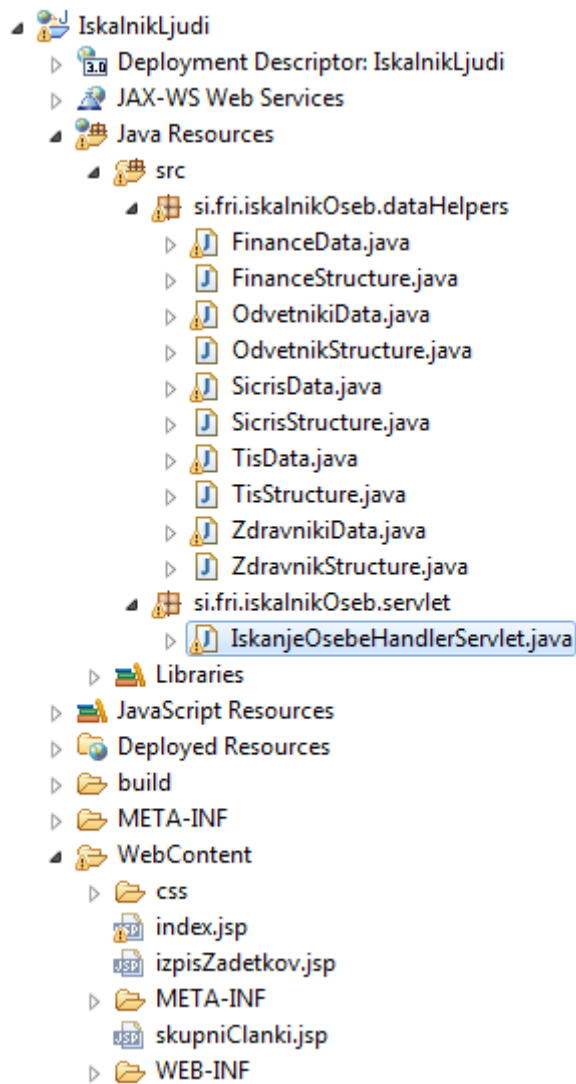
Apache POI je knjižnica Java za branje in pisanje Microsoftovih dokumentov, kot so Word, PowerPoint in Excel [37].

5.2 Izvorna koda aplikacije

Izvorna koda aplikacije je prosto dostopna na repozitoriju Git na naslovu <https://github.com/mznide/IskalnikLjudi>.

5.3 Zgradba aplikacije

Aplikacija je sestavljena iz različnih sklopov datotek: datoteke JSP skrbijo za strukturo strani; Servlet se sproži ob začetku iskanja in skrbi za logiko aplikacije; datoteke, ki so v paketu dataHelpers, skrbijo za povezavo do različnih strani in pridobitev podatkov; datoteke CSS skrbijo za vizualno podobo aplikacije. Zgradbo projekta prikazuje slika 12.



Slika 12: Zgradba aplikacije

5.4 Uporabniški vmesnik

Za izdelavo videza aplikacije smo uporabili kaskadne stilske predloge CSS. Z njimi smo poskrbeli za pozicijo elementov, barvo ozadja, obliko pisave ter videz vnosnega polja in gumba. Sledili smo modernim smernicam in dodali samo tiste gradnike, ki jih za aplikacijo res potrebujemo.

Slika 13 prikazuje osnovno stran naše aplikacije, ki se prikaže ob zagonu. Ponuja nam možnost vpisa imena in priimka ter začetka iskanja informacij o osebi. Zraven ponuja še nekaj primerov povezav, ki sprožijo iskanje za navedeno osebo.

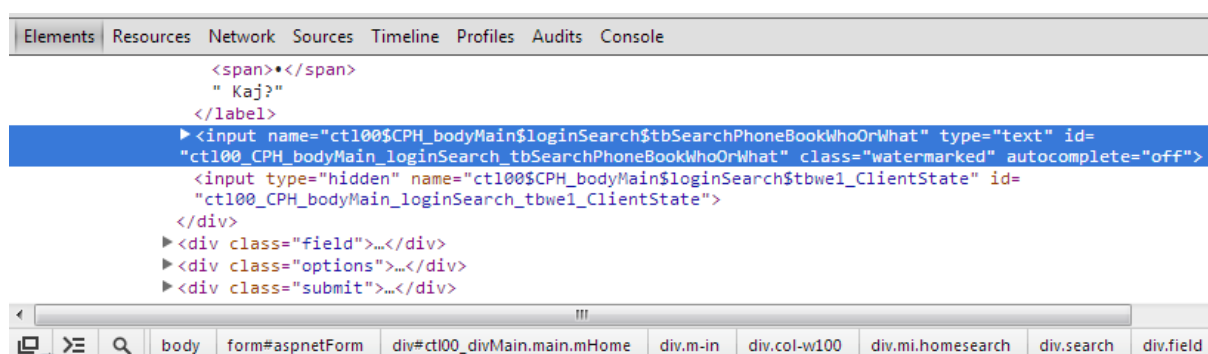


Slika 13: Osnovna stran aplikacije

5.5 Luščenje iz Telefonskega imenika Slovenije

Telefonski imenik Slovenije na strani ponuja vgrajen iskalnik, ki omogoča iskanje po osebah. Odločili smo se, da bomo za luščenje podatkov uporabili knjižnico HtmlUnit in si tako močno olajšali delo. Namesto ročnega pošiljanja zahtev GET in POST bomo simulirali uporabnika in njegove klike. Za stran bomo videti kot navaden uporabnik spleta, in ne kot spletni robot.

Najprej bomo poslali zahtevo za pridobitev glavne strani <http://www.itis.si>. Za iskanje osebe bo treba izpolniti iskalno formo in pritisniti gumb najdi. Potrebujemo element textbox html za vpis imena in priimka ter gumb najdi. Želeni elementi imajo imena in bomo tako najlažje dostopali do njih. Najprej s pomočjo orodja Chrome developers tools ugotovimo imena ustreznih elementov.



Slika 14: Chrome developer tools

Nato pošljemo ustrezne zahteve za pridobitev zelenih elementov HTML. Elementu textbox spremenimo vrednost v ime naše osebe. Na elementu, ki predstavlja gumb, pa izvedemo akcijo, ki simulira klik.

```
final WebClient webClient = new WebClient(BrowserVersion.FIREFOX_10);
webClient.getOptions().setJavaScriptEnabled(false);

// Posljemo zahtevo za osnovno stran
final HtmlPage page = webClient.getPage("http://www.itis.si/");

// Posljemo zahteve za pridobitev ustreznih html elementov
final HtmlForm form = page.getFormByName("aspnetForm");

final HtmlSubmitInput button = form
    .getInputByName("ctl00$CPH_bodyMain$loginSearch$btnSearchPhoneBook");

final HtmlTextInput textFieldIme = form
    .getInputByName("ctl00$CPH_bodyMain$loginSearch$tbSearchPhoneBookWhoOrWhat");
textFieldIme.setValueAttribute(name);

HtmlPage pageIskanje = button.click();
```

Stran, ki se prikaže, vsebuje naše rezultate. V primeru, da je v Sloveniji več ljudi z istim imenom in priimkom, je lahko teh rezultatov več. Želeni elementi nimajo imen in bomo morali do njih dostopati s pomočjo XPatha. Napišemo ustrezne poizvedbe XPath in tako za zadetke pridobimo ime, ulico, pošto in telefonsko številko.

```
List<HtmlElement> imenaOseb = (List<HtmlElement>) pageOsebe
    .getByXPath("//div[@class=\"r-in\"]//h3//a");
List<HtmlElement> ulicaOseb = (List<HtmlElement>) pageOsebe
    .getByXPath("//div[@class=\"r-in\"]//p[@class=\"address\"]//a//span[@class=\"street\"]");
List<HtmlElement> postaOseb = (List<HtmlElement>) pageOsebe
    .getByXPath("//div[@class=\"r-in\"]//p[@class=\"address\"]//a//span[@class=\"zip\"]");
List<HtmlElement> telOseb = (List<HtmlElement>) pageOsebe
    .getByXPath("//div[@class=\"r-in\"]//p//span[@class=\"phone\" or
    @class=\"gsm\"]//a");
```

5.6 Luščenje podatkov iz ZZZS

Pri luščenju podatkov iz ZZZS smo naleteli na tri Excelove datoteke. Gre za strukturirane podatke, tako da bo luščenje nekoliko bolj preprosto. Spodnja slika prikazuje zgradbo datotek.

Šifra OE	Naziv OE oz. izpostave	Šifra izvajalca	Naziv izvajalca	Šifra zdravnika	Priimek in ime zdravnika	ZZS dejavnost	Število timov	Število opredeljenih v letih				
								0	1 do 6	7 do 18	19 do 49	50 do 64
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	71	HRUŠOVAR MARJAN	302001	1,00	0	-	11	645	547
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	28	ČAJAVEC RUDI	302001	0,24	0	-	3	430	326
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	3269	VUKOREPA BUHANEC BIJANA	302001	1,00	0	-	-	989	314
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	3979	SELIČ AMON MARGARETA	302001	1,00	0	1	10	604	508
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	4054	ŽMAVC ANDREJ	302001	0,60	0	-	-	252	208
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	4625	MALINAR RAJKA	302001	0,50	0	-	1	69	58
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	5365	JUNEŽ MARIJA	302001	1,00	0	-	-	546	608
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	4988	OŠEP STANKO	302001	1,00	0	1	13	734	663
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	5119	ŽITKO MASTNAK MILJANA	302001	0,20	0	-	1	262	221
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	5158	POVH JESENŠEK ERIKA	302001	0,48	0	-	-	#####	195
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	5196	GOVC ERŽEN JANA	302001	1,00	7	53	297	#####	324
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	6277	RAK ALENKA	302001	1,00	3	52	198	934	429
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	6317	ĐORDEVIĆ STEVAN	302001	1,00	1	52	224	#####	595
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	6107	BIZJAK SLEMENŠEK BREDA	302001	0,20	0	-	-	123	48
2E+05	OE CELJE	2131	ZD ZDRAVSTVENI DOM CELJE	6239	KAJBA VENINŠEK SIMONA	302001	1,00	0	-	-	642	675

Slika 15: Excelova datoteka iz ZZS

Knjižica Apache-POI nam omogoča dostop do Microsoftovih dokumentov, med katere spadajo tudi naše datoteke. Najprej je treba narediti instanco objekta `HSSFWorkbook`, ki omogoča dostop do datoteke. Nato izberemo prvo stran dokumenta in se s pomočjo iteratorjev sprehodimo po njej. Zanimali nas bodo naziv izpostave, naziv izvajalca ter ime in priimek. Če se ime iskane osebe ujema s prebranim elementom, imamo zadetek [37].

```

// Dobimo workbook instanco naše datoteke
HSSFWorkbook workbook = new HSSFWorkbook(file1);

// Dobimo prvo stran datoteke
HSSFSheet sheet = workbook.getSheetAt(0);

// Iterator po vrsticah na trenutni strani
Iterator<Row> rowIterator = sheet.iterator();

int i = 0;
while (rowIterator.hasNext()) {
    i++;
    Row row = rowIterator.next();
    Iterator<Cell> cellIterator = row.cellIterator();
    if (i >= 8) {
        String nazivOE = "";
        String nazivIzvajalca = "";
        int j = 0;
        while (cellIterator.hasNext()) {
            j++;
            Cell cell = cellIterator.next();
            if (j == 2)
                nazivOE = cell.getStringCellValue();
            if (j == 4)
                nazivIzvajalca = cell.getStringCellValue();
            if (j == 6 &&
                (name.toUpperCase().equals(cell.getStringCellValue())))
                return new ZdravnikStructure(name, nazivOE,
                                                nazivIzvajalca);
        }
    }
}

```

5.7 Luščenje podatkov iz časopisa Finance

Spletna stran [finance.si](http://www.finance.si) vse osebe, ki nastopajo v člankih, označi oz. taga in naredi povezavo do strani, kjer se da o osebi izvedeti več. Vse take strani oseb in podjetij je možno najti na <http://www.finance.si/leksikon>, kjer je možno brskati po abecedi ali s pomočjo iskalnika.

Leksikon: zbirka imen, omenjenih v člankih

Uporabite iskalno škatlico spodaj

Podjetja ali institucije, omenjene v člankih, najdete poleg obarvanih povezav v besedilu tudi v škatlici "Omenjeni", ki se ob branju pojavi na desnem robu zaslona.

Čeprav so v objavljenih člankih posebej označene tudi osebe, jih v tej referenčni zbirki ni mogoče poiskati. Razlog za to je trenutno veljavno tolmačenje zakona o varstvu osebnih podatkov.

Poišči...

Ime ali del imena:

Poišči

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z Č Š Ž 0 1 2 3 4 5 6 7 8 9

Slika 16: Iskalnik strani finance.si

Za pridobitev člankov, v katerih se pojavlja naša oseba, bomo uporabili iskalnik. S pomočjo HtmlUnita bomo vnesli ime iskane osebe in klicali funkcijo, ki simulira pritisk gumba »Poišči«. V večini primerov bomo dobili en rezultat, na katerega zopet simuliramo klik. Če je rezultatov več, bomo kliknili na prvega, saj je ta najbolj aktualen. Dobili bomo stran s povezavami do desetih člankov, v katerih je bila oseba nazadnje omenjena. Te s pomočjo XPatha izluščimo in vrnemo kot rezultat.

5.8 Implementacija funkcije poznanstva

Poleg luščenja člankov, v katerih je omenjena iskana oseba, pa želimo stran finance.si uporabiti tudi za implementacijo funkcije, ki bo iskala poznanstva osebe. Delovala bo tako, da se bomo sprehodili po člankih, v katerih je oseba omenjena, in pogledali, s kom se najbolj pogosto pojavlja skupaj. Te osebe bomo nato uredili po pogostosti in najbolj pogostih deset prikazali kot rezultat. Zadeve se bomo zopet lotili s knjižnico HtmlUnit.

Postopek, ki ga aplikacija uporablja za pridobitev povezav, je naslednji:

1. Pošlji poizvedbo za stran <http://www.finance.si/leksikon>.
2. V zgornjo formo naj se vpiše ime iskane osebe in pritisne gumb »Poišči«.
3. Kliknemo prvi in po navadi edini rezultat, ki ga vrne poizvedba.
4. Dobili bomo seznam desetih člankov, v katerih se je oseba nazadnje pojavila. Tokrat nas bo zanimalo več člankov, zato bo treba klikniti povezavo »več«.
5. Dobimo seznam člankov, v katerih je oseba omenjena. Na strani je lahko največ trideset člankov, na naslednjo stran pa se je možno pomakniti s klikom na »starejše«.
6. S pomočjo XPatha izluščimo povezave do člankov in se pomaknemo na naslednjo stran. Skupaj izluščimo do sto petdeset povezav do člankov.

7. Pošljemo poizvedbo za vsak članek in izluščimo osebe in pravne subjekte, ki se pojavljajo v članku.
8. Seznam oseb in pravnih subjektov uredimo po pogostosti ter ga vrnemo kot rezultat. Najpogostejših deset prikažemo kot rezultat.

5.9 Integracija podatkov

Pri luščenju iz SICRIS-a gre za podobno luščenje kot pri luščenju TIS-a in ga zato ne opisujemo posebej. Podatke, ki smo jih izluščili iz različnih virov, pa je treba združiti in prikazati na enem mestu. V servletu, ki se ga kliče ob začetku iskanja, je prikazano luščenje podatkov iz petih virov in na koncu preusmeritev na prikaz rezultatov.

```
protected void doPost(HttpServletRequest request,
    HttpServletResponse response) throws ServletException,
IOException {
    request.setCharacterEncoding("UTF-8");
    String nameOfPerson = request.getParameter("nameOfPerson");
    try {
        ZdravnikStructure zdravnik = ZdravnikiData
            .getZdravnikData(
                request.getServletContext()
                    .getRealPath("/WEB-INF/"), nameOfPerson);
        SicrisStructure[] osebeSicris = SicrisData
            .getSicrisData(nameOfPerson);
        TisStructure[] osebeTis = TisData.submittingForm(nameOfPerson);
        OdvetnikStructure odvetnik = OdvetnikiData
            .getOdvetnikData(nameOfPerson);
        ArrayList<FinanceStructure> povezave = FinanceData
            .getPovezaveOsebe(nameOfPerson);
        String[] clanki = FinanceData.getCLankeOsebe(nameOfPerson);

        request.getSession().setAttribute("iskanaOseba", nameOfPerson);
        request.getSession().setAttribute("osebeTis", osebeTis);
        request.getSession().setAttribute("osebeSicris", osebeSicris);
        request.getSession().setAttribute("zdravnik", zdravnik);
        request.getSession().setAttribute("odvetnik", odvetnik);
        request.getSession().setAttribute("povezave", povezave);
        request.getSession().setAttribute("clanki", clanki);

        response.sendRedirect("izpisZadetkov.jsp");

    } catch (Exception e) {
        // TODO Auto-generated catch block
        e.printStackTrace();
    }
}
```

5.10 Prilagoditev za šumnike

Ker gre za slovensko aplikacijo in se bodo v iskalnih poizvedbah in rezultatih pogosto pojavljali šumniki, je pomembno, da ti pravilno delujejo. Uporabili bomo nabor znakov UTF-8, ki podpira slovensko abecedo.

Kodiranje strani JSP

Za pravilen prikaz datotek JSP je treba na vrhu vsakega dokumenta JSP dodati spodnjo direktivo. S tem bomo dali brskalniku vedeti, da je stran kodirana z naborom UTF-8, ki vsebuje tudi šumnike.

```
<%@ page language="java" contentType="text/html; charset=UTF-8"
    pageEncoding="UTF-8"%>
```

Metaoznake html

Nastaviti je treba tudi metaoznake, saj želimo, da tudi te delujejo pravilno.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>
<head>
<link rel="stylesheet" href="css/index.css" />
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
<title>Iskalnik ljudi</title>
</head>
```

Konfiguracijska datoteka standalone.xml

Če želimo, da naš strežnik pravilno procesira zahteve GET in POST, je treba strežnik nastaviti za kodiranje v naboru UTF-8. V primeru našega strežnika JBoss AS 7.1 to storimo tako, da v konfiguracijsko datoteko standalone.xml dodamo naslednjo kodo.

```
<system-properties>
    <property name="org.apache.catalina.connector.URI_ENCODING" value="UTF-8"/>
    <property
name="org.apache.catalina.connector.USE_BODY_ENCODING_FOR_QUERY_STRING"
value="true"/>
</system-properties>
```

Servlet GET in POST

Pred branjem parametra je treba nastaviti tudi kodiranje naše zahteve.

```
request.setCharacterEncoding("UTF-8");
String nameOfPerson = request.getParameter("name");
```

5.11 Prikaz uporabe in primeri poizvedb

Miro Senica

Isci osebo

Rezultati za osebo Miro Senica:

Telefonski imenik Slovenije
(Vir: Telefonski imenik Slovenije)
Ime: SENICA MIROSLAV
Telefon: 03 579 10 34
Naslov: Planina pri Sevnici 14,
Posta: 3225 Planina pri Sevnici
[Več podatkov](#)

Raziskovalec
(Vir: Sliris.si)

Zdravnik
(Vir: ZZZS)

Odvetnik
(Vir: Odvetniška zbornica slovenije)
Ime: Miro Senica
Službeni telefon: 01 252 80 00
Naslov: Barjanska cesta 3
Kraj: Ljubljana

Seznam clankov v katerih je oseba omenjena
(Vir: Finance.si)
[Mišič: V Luki je menda prisen tudi Senica](#)
[Senici naj bi blokirali premoženje](#)
[Se Rakovec lahko spet povzpne na vrh Triglava?](#)
[\(hiti tedna\) Katarina Kresal za Mercator rešuje potrošniške spore](#)
[Katarina Kresal za Mercator rešuje potrošniške spore](#)
[Nove preiskave kreativnih davčnih utajevalcev](#)
[Kmetija Bužekijan korak bliže rušitvi](#)
[Mesarsko klanje za ostaline Vegrada](#)
[Igorju J. Pogačarju presihajo proračunski prihodki](#)
[Mikoličeva zavrača očitke glede razrešitve](#)

Seznam povezav
(Vir: Finance.si)
[Katarina Kresal\(54 \)](#)
[Zoran Jankovič\(15 \)](#)
[Borut Pahor\(14 \)](#)
[Vegrad\(13 \)](#)
[NLB\(11 \)](#)
[Petroli\(11 \)](#)
[Janez Janša\(10 \)](#)
[Igor Jurij Pogačar\(9 \)](#)



Slika 17: Primer iskanja odvetnika

Živo Bobič

Išci osebo

Rezultati za osebo Živo Bobič:

Telefonski imenik Slovenije

(Vir: [Telefonski imenik Slovenije](#))

Ime: BOBIČ ŽIVO
Telefon: 01 364 43 20
Naslov: Visoko 150, Visoko
Posta: 1292 Ig

[Več podatkov](#)

Raziskovalec

(Vir: [Slois.si](#))

Zdravnik

(Vir: [ZZS](#))

Ime: Živo Bobič
OE: ZOB BABIT D.O.O.
Naziv izvajalca: ZOB BABIT D.O.O.

Odvetnik

(Vir: [Odvetniška zbornica slovenije](#))

Seznam clankov v katerih je oseba omenjena

(Vir: [Finance.si](#))

[Največji v zdravstveni dejavnosti](#)
[Koliko zaslužijo zasebni zobozdravniki](#)
[Koliko zaslužijo zasebni zobozdravniki](#)

Seznam povezav

(Vir: [Finance.si](#))



Slika 18: Primer iskanja zdravnika

Janez Balažic

Išci osebo

Rezultati za osebo Janez Balažic:

Telefonski imenik Slovenije

(Vir: [Telefonski imenik Slovenije](#))

Ime: BALAŽIC JANEZ
Telefon: 01 723 14 60
Naslov: Dole pri Krašcah 22, Dole pri Krašcah
Posta: 1251 Moravče

Ime: BALAŽIC JANEZ
Telefon: 031 371 898
Naslov: Stritarjeva ulica 22, Vir
Posta: 1230 Domžale

Ime: BALAŽIC JANEZ
Telefon: 02 541 20 04
Naslov: Žitna ulica 3,
Posta: 9231 Beltinci

[Več podatkov](#)

Raziskovalec

(Vir: [Slois.si](#))

Ime: Dr. Balažic Janez
Telefon: (02) 521 11 55
Fax številka: (02) 521 11 55
Email: Janez.Balazic@uni-mb.si
Website:

Seznam clankov v katerih je oseba omenjena

(Vir: [Finance.si](#))

[Harmonija med dediščino preteklosti in sedanjostjo](#)
[Odklop](#)
[Enkrat sem, drugič tja](#)

Seznam povezav

(Vir: [Finance.si](#))



Slika 19: Primer iskanja raziskovalca

Ime in priimek

Isci osebo

Prikaz skupnih člankov oseb

Alenka Bratušek in **Karel Lipnik**:

[Bratušek: Brez dviga DDV ne bo rasti](#)

['Najprej treba rešiti bančni krč, potem zadolženost gospodarstva'](#)

[Bratuškovo skrbijo izjave ministrov in poslancev koalicije](#)

[Lipnik: V Sloveniji ne pride v poštev ne dvig davkov in ne spodbujanje porabe](#)

[Bratušek: Zakaj moramo vedno čakati signal od zunaj in delati po navodilih?](#)

['Janševi varčevalni ukrepi bodo imeli negativni vpliv na BDP'](#)

[Ali vladno varčevanje zavira gospodarstvo?](#)

[Pojdi nazaj](#)

Slika 20: Skupni članki dveh oseb

6. Sklepne ugotovitve

Ugotovili smo, da je slovenski splet razvit in ponuja veliko različnih storitev. Nekatere izmed njih so lahko zanimiv in koristen vir podatkov o osebah. Tudi v Sloveniji je internet zelo razširjen in je teh virov ogromno. Ponujajo nam najrazličnejše informacije, kot so telefonska številka, domači naslov, poklicna dejavnost. Prisotni so različni članki in spletne objave na forumih, socialna omrežja pa nam razkrijejo še veliko več. S časom lahko pričakujemo, da se bo obseg spletnih storitev še povečal in bo teh virov vedno več.

Jasno je postalo, da je treba biti na spletu previden; kar objavljamo na spletu, je preprosto najti. Uporabniki ali spletni roboti lahko podatke shranijo in imajo tako dostop do njih tudi po našem izbrisu. Kljub temu da nam to kdaj pride prav, pa to zmanjšuje zasebnost in ljudje lahko o nas zvedo več, kot bi si sami morda želeli.

Implementacija iskalnika nam je uspela. Povezali smo pet različnih virov in jih združili med seboj. Naš iskalnik tako omogoča pridobitev naslova, telefonske številke ter v določenih primerih tudi poklica osebe. Najdemo lahko tudi članke, v katerih se pojavlja iskana oseba, ter s kom se v člankih pogosto pojavlja, na podlagi česar lahko sklepamo o določenih lastnostih iskane osebe. Dodatno nam iskalnik omogoča prikaz skupnih člankov dveh oseb.

Pri luščenju in integraciji podatkov smo se morali držati zakonodaje in spoštovati želje lastnikov strani, do katerih smo dostopali, kar je prineslo nekaj omejitev. Naša aplikacija se lahko uporablja kot pripomoček pri iskanju, ne moremo pa je javno ponuditi na spletu kot storitev. Podatki se ne shranjujejo lokalno, ampak se med uporabo dostopa do različnih spletnih virov. Iskalnik ni namenjen komercialni uporabi.

Aplikacija je v fazi prototipa in lahko služi kot osnova za nadaljnji razvoj. Možnih izboljšav je veliko. V drugem poglavju smo izdelali konceptualni model slovenskih spletnih virov, ki prikazuje, kako bi ta razvoj lahko potekal. Izboljšave so predvsem v dodajanju novih virov in izboljševanju integracije. Aplikacija je izdelana na način, ki tako dodajanje omogoča. Z novimi viri bi se povečala količina informacij in bi tako lahko o osebi izvedeli več. Z vključitvijo katerega izmed socialnih omrežij in različnih forumov bi aplikacija postala bolj usmerjena na osebne podatke o osebi, vključitev poslovnih imenikov pa bi aplikacijo naredilo uporabno za delodajalce. Razvijalec bi lahko spremenil namen aplikacije in izbral enega izmed opisanih možnih scenarijev integracij ali pa se odločil za novega. S pravimi izbranimi viri in ustrezno integracijo bi si olajšal delo.

Literatura

- [1] (10. 9. 2013) „Inventing the Web: Tim Berners-Lee 1990 Christmas Baby « Seeing the Picture,“ Dostopno na: <http://blog.lib.uiowa.edu/hardinmd/2010/11/24/inventing-the-web-tim-berners-lees-1990-christmas-baby/>.
- [2] B. Sanou, „The World in 2013 ICT Facts and Figures“.
- [3] M. Turland, php|architect's Guide to Web Scraping with PHP, Canada, 2010.
- [4] (10. 9. 2013) „Web Service Architecture,“ [Elektronski]. Dostopno na: <http://www.w3.org/TR/ws-arch/#whatis>.
- [5] (10. 9. 2013) „XML Path Language (XPath) 2.0 (Second Edition),“ [Elektronski]. Dostopno na: <http://www.w3.org/TR/xpath20/#id-introduction>.
- [6] (10. 9. 2013) „Regular Expressions,“ [Elektronski]. Dostopno na: http://pubs.opengroup.org/onlinepubs/009695399/basedefs/xbd_chap09.html.
- [7] (10. 9. 2013) „W3C Document Object Model,“ [Elektronski]. Dostopno na: <http://www.w3.org/DOM/>.
- [8] (10. 9. 2013) „The W3C Document Object Model,“ [Elektronski]. Dostopno na: <http://www.w3.org/2002/07/26-dom-article.html>.
- [9] (10. 9. 2013) „Selenium - Web Browser Automation,“ [Elektronski]. Dostopno na: <http://docs.seleniumhq.org/>.
- [10] (10. 9. 2013) „HtmlUnit - Welcome to HtmlUnit,“ [Elektronski]. Dostopno na: <http://htmlunit.sourceforge.net/>.
- [11] (10. 9. 2013) „Sahi Web Test Automation Tool,“ [Elektronski]. Dostopno na: <http://sahi.co.in/>.
- [12] (10. 9. 2013) „Uradni list Republike Slovenije,“ [Elektronski]. Dostopno na: <http://www.uradni-list.si/1/content?id=86833>.
- [13] (10. 9. 2013) P. Zdešar, „Statistični urad RS - Uporaba informacijsko - komunikacijske tehnologije v gospodinjstvih in pri posameznikih, Slovenija, 2012 - končni podatki,“ [Elektronski]. Dostopno na: http://www.stat.si/novica_prikazi.aspx?ID=5037.

- [14] (10. 9. 2013) „TIS - Telefonski imenik Slovenije,“ [Elektronski]. Dostopno na: <http://www.itis.si>.
- [15] (10. 9. 2013) „Poslovni imenik bizi.si,“ [Elektronski]. Dostopno na: <http://www.bizi.si/>.
- [16] (10. 9. 2013) „Finance.si,“ [Elektronski]. Dostopno na: <http://www.finance.si/>.
- [17] (10. 9. 2013) „O mreži COBISS.net,“ [Elektronski]. Dostopno na: http://www.cobiss.net/o_mrezi_COBISS-Net.htm.
- [18] (10. 9. 2013) „SICRIS,“ [Elektronski]. Dostopno na: <http://www.sicris.si/about/cris.aspx?lang=slv>.
- [19] (10. 9. 2013) „Slovenska Wikipedia praznuje prvih deset let obstoja,“ [Elektronski]. Dostopno na: <http://www.delo.si/druzba/panorama/slovenska-wikipedia-praznuje-prvih-deset-let-obstoja.html>.
- [20] (10. 9. 2013) I. Križman, „Statistični urad Republike Slovenije - Državna statistika,“ [Elektronski]. Dostopno na: http://www.stat.si/drz_stat.asp.
- [21] (10. 9. 2013) „Beseda o nas,“ [Elektronski]. Dostopno na: <http://www.zzzs.si/zzzs/internet/zzzs.nsf/o/EE17AD1D28A89B2DC1256E8600348651#>.
- [22] (10. 9. 2013) „OZS | Informacije javnega značaja,“ [Elektronski]. Dostopno na: <http://www.odv-zb.si/predstavitev/informacije-javnega-znacaja>.
- [23] (10. 9. 2013) „Seznam neplačnikov | Davčna uprava RS,“ [Elektronski]. Dostopno na: http://seznam.gov.si/DURS/main_neplacniki.html.
- [24] (10. 9. 2013) „RIS - Raba Interneta v Sloveniji,“ [Elektronski]. Dostopno na: http://www.ris.org/db/27/12535/Raziskave/Na_Facebooku_skoraj_750000_slovenskih_uporabnikov/.
- [25] (10. 9. 2013) „Discover. Monitor. Protect. | 123people.com,“ [Elektronski]. Dostopno na: <http://www.123people.com/>.
- [26] (10. 9. 2013) „All about us | 123people.com,“ [Elektronski]. Dostopno na: <https://www.123people.com/page/about>.

- [27] (10. 9. 2013) „Welcome to Deep Web - Pipl - People Search,“ [Elektronski]. Dostopno na: <https://pipl.com/help/deep-web/>.
- [28] (10. 9. 2013) „ZabaSearch | Free People Search | Search Menu,“ [Elektronski]. Dostopno na: http://www.zabasearch.com/search_menu/.
- [29] (10. 9. 2013) „Supervizor,“ [Elektronski]. Dostopno na: <http://supervizor.kpk-rs.si/oprojektu/>.
- [30] (10. 9. 2013) „Kdo kliče? - Android Apps on Google Play,“ [Elektronski]. Dostopno na: <https://play.google.com/store/apps/details?id=com.adaffix.market.si.android>.
- [31] (10. 9. 2013) „About the Eclipse Foundation,“ [Elektronski]. Dostopno na: [http://www.eclipse.org/org/..](http://www.eclipse.org/org/)
- [32] (10. 9. 2013) „JBoss Application Server 7 - JBoss Community,“ [Elektronski]. Dostopno na: <https://www.jboss.org/jbossas/>.
- [33] (10. 9. 2013) S. D. James Clark, „XML Path Language (XPath),“ [Elektronski]. Dostopno na: <http://www.w3.org/TR/xpath/>.
- [34] (10. 9. 2013) A. L. H. I. J. Dave Raggett, „HTML 4.01 Specification,“ [Elektronski]. Dostopno na: <http://www.w3.org/TR/html401/>.
- [35] (10. 9. 2013) W3C, „HTML & CSS - W3C,“ [Elektronski]. Dostopno na: <http://www.w3.org/standards/webdesign/htmlcss#whatcss>.
- [36] (10. 9. 2013) „Chrome DevTools - Google Developers,“ [Elektronski]. Dostopno na: <https://developers.google.com/chrome-developer-tools/>.
- [37] (10. 9. 2013) „Apache POI - the Java API for Microsoft Documents,“ [Elektronski]. Dostopno na: <http://poi.apache.org/>.