Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Sašo Marić

Vmesnik za dostop do portala odprtih podatkov Slovenije

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: doc. dr. Tomaž Curk

Ljubljana, 2018

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Nacionalni portal Odprtih podatkov Slovenije (OPSI) omogoča dostop do bogatega vira nekaj tisoč zbirk podatkov o štirinajstih področjih delovanja javnega sektorja. Vsaka zbirka je opisana z metapodatki o vsebini ter pogojih dostopa in ponovne uporabe podatkov. V okviru naloge razvijte programski vmesnik, ki bo omogočal preprost dostop do podatkov. Vmesnik naj prepozna osnovne podatkovne tipe, kot sta tabela primerov, besedilo, slika. Podatke naj preoblikuje v obliko, ki jo podpira programski paket za podatkovno rudarjenje Orange.

Kazalo

Povzetek

Abstract

1 Uvod **1**

1.1 Portal OPSI 1

1.2 Orange 3

1.3 Motivacija 3

1.4 Cilji 3

2 Pregled področja in metode **5**

2.1 Obstoječi portali odprtih podatkov 5

2.2 Obstoječe rešitve za pridobivanje podatkov iz spleta 6

2.3 Uporabljene metode pri izdelavi vmesnika 7

3 Razvoj **11**

3.1 Arhitektura vmesnika 12

3.2 Postavitev okolja 17

3.3 Razvoj vmesnika 18

3.4 Algoritem določanja tipa atributov 19

3.5 Testiranje in težave 20

4 Rezultati **23**

4.1 Primer uvoza podatkov iz vmesnika v program Orange 23

4.2 Uspešnost detekcije in pretvorbe podatkov 28

5 Sklepne ugotovitve **35**

Literatura **37**

Seznam uporabljenih kratic

kratica	angleško	slovensko
API	Application programming interface	programski vmesnik
Cron	Time-based job scheduler	časovno nastavljeno periodično opravilo
GUI	Graphical user interface	grafični uporabniški vmesnik
HTML	Hypertext Markup Language	označevalni jezik za oblikovanje večpredstavnostnih dokumentov
HTTP	Hypertext Transfer Protocol	protokol za izmenjavo hipertesksta ter grafičnih, zvočnih in drugih večpredstavnostnih vsebin na spletu
JSON	JavaScript Object Notation	objektna notacija v JavaScript
OPSI	Open Data of Slovenia	Odprti podatki Slovenije
REST	Representational state transfer	arhitekturni stil za komunikacijo med spletnimi storitvami

Povzetek

Naslov: Vmesnik za dostop do portala odprtih podatkov Slovenije

Avtor: Sašo Marić

Portal Odprti Podatki Slovenije (OPSI) je vzpostavilo Ministrstvo za javno upravo s ciljem zagotoviti celostni popis podatkovnih zbirk, ki jih vodijo organi javnega sektorja, ter omogočiti objavo zbirk v obliki odprtih podatkov. Vsi podatki na portalu so podatki, evidence in zbirke, ki nastajajo pri delu organov javnega sektorja in so prosto dostopni.

V diplomski nalogi smo obdelali dostopne podatke in jih pripravili v obliki za uvoz v program za strojno učenje in vizualizacijo podatkov Orange. Prikazali smo tudi statistiko metapodatkov na portalu OPSI. Vmesnik uspešno pretvori 788 datotek od 19565. Največ virov datotek in sicer 375 izhaja iz področja Vlada in javni sektor. V datotekah s preglednicami je 1038 stolpcev oz. spremenljivk, ki jih delimo v štiri skupine: diskretne (464), nizi (257), zvezne (180) in časovne znamke (137). Večina datotek (13765 datotek), ki jih vmesnik ne more pretvoriti samodejno, je formata html.

S pomočjo vmesnika lahko portalu OPSI sporočamo pripombe ali izboljšave glede podatkov in javljamo nepravilnosti na portalu in podatkih. Podatke, ki jih razviti vmesnik samodejno pretvori, lahko uporabimo v programskem orodju Orange in v njih odkrivamo zakonitosti in zanimive vzorce.

Ključne besede: odprti podatki, obdelava podatkov, pridobivanje spletnih podatkov.

Abstract

Title: Interface for accessing the portal Open data of Slovenia

Author: Sašo Marić

The portal Odprti Podatki Slovenije (OPSI) was established by the Ministry of Public Administration with the aim to provide an integrated listing of databases managed by the bodies of the public sector and to allow easy publication of data collections in form of open data. All the data provided on the portal are records and collections, which are created by the public sector bodies and are freely accessible.

We have preprocessed the freely accessible data and formatted it into the format for Orange data mining, which is a program for machine learning and data visualization. We provide statistics on the available data in the OPSI portal. The interface automatically transforms 788 files out of 19565. The largest source of files, namely 375, is the area of the Government and the public sector. The interface automatically converts data in 1038 columns total, which are grouped into four categories: discrete (464), string (257), continuous (180) and time (137). Most of the files (13765 files) that the interface cannot transform automatically are in the html format.

With the interface we can provide comments or improvements regarding the published data and report any irregularities of the portal and data. The data provided by the interface can be mined for interesting patterns using the Orange data mining software.

Keywords: open data, data processing, web scraping.

Poglavje 1

Uvod

Odprti podatki temeljijo na dejstvu, da so prosto dostopni. Njihov namen je zagotavljanje transparentnosti delovanja javnih služb ter razpoložljivost in ponovna uporaba podatkov z namenom ustvarjanja novega znanja, npr., novih aplikacij, brez avtorskih omejitev [1]. V veliki večini so podatki dostopni preko spletnih vmesnikov predvsem zaradi lažjega iskanja. Pojem odprti podatki je relativno nov in hitro pridobiva na priljubljenosti predvsem zaradi hitrega vzpona svetovnega spleta [2]. Pobuda odprtih podatkov prihaja predvsem zaradi evropskega zakona Direktiva o ponovni uporabi podatkov javnega sektorja. V Sloveniji je bil kasneje sprejet tudi Zakon o dostopu do informacij javnega značaja [3].

1.1 Portal OPSI

OPSI je nacionalni portal odprtih podatkov vzpostavljen na podlagi EU Direktive o ponovni uporabi podatkov javnega sektorja in zakonodaje o dostopu do informacij javnega značaja [4]. Portal OPSI je namenjen objavam odprtih podatkov. Na njem objavljajo vsi organi javne uprave kot tudi uporabniki podatkov [5]. Vzpostavljen je bil leta 2016 z namenom dostopnosti in ponovne uporabe podatkov javnega značaja. Ciljna publika portala je širša javnost, zato so vsi podatki na portalu brezplačni in namenjeni nadaljnji ob-

delavi tudi za komercialne namene [6]. OPSI trenutno šteje 3748 podatkovnih zbirk iz 14 različnih področij:

- prebivalstvo in družba,
- pravosodje, pravni sistem in javna varnost,
- vlada in javni sektor,
- izobraževanje, kultura in šport,
- sociala in zaposlovanje,
- zdravje,
- okolje in prostor,
- promet in infrastruktura,
- kmetijstvo, ribištvo, gozdarstvo in prehrana,
- finance in davki,
- gospodarstvo,
- energetika,
- znanost in tehnologija,
- mednarodne zadeve.

Podatki so shranjeni v različnih formatih podatkovnih zbirk: csv, docx, txt, html in drugi. Poleg podatkov portal OPSI ponuja tudi vmesnik API za metapodatke, s katerim lahko na enostaven način izluščimo datoteke posameznih podatkovnih virov, ki nam jih portal ponuja. Vmesnik API omogoča beleženje sprememb podatkov na portalu.

V okviru diplomskega dela smo pripravili uporabniku prijazen vmesnik, preko katerega lahko uporabnik hitro in enostavno dostopa do želenih podatkov posameznega področja. Vmesnik deluje preko klicev API. Iz podatkovne baze, z vnaprej pripravljenimi vsebinami, uporabniku omogoča dostop do datoteke ob kliku na željen podatkovni vir. Podatki so pripravljeni v obliki, ki je primerna za uvoz v program za podatkovno rudarjenje Orange, kjer lahko podatke prikažemo ali dodatno obdelamo. Pripravili smo tudi statistiko podatkov ter metapodatkov, ki jih je bilo moč preoblikovati in shraniti v podatkovno bazo. Uporabnik tako dobi hiter vpogled v podatke shranjene v OPSI ter njihovo uporabnost.

1.2 Orange

Orange je odprtokodno orodje za vizualizacijo ter rudarjenje podatkov in strojno učenje. Orodje omogoča vizualno programiranje, kjer elemente upravljamo grafično. Orange razvijajo na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Začetki razvoja orodja segajo v leto 1997 [7]. Orodje je zelo enostavno za uporabo. V mnogih ustanovah ga uporabljajo kot učni pripomoček pri učenju osnov obdelave podatkov. Sestavljen je iz več komponent imenovanih slikovni gradniki, s katerimi lahko podatke prikažemo, procesiramo, odkrivamo vzorce v podatkih, gradimo in vrednotimo napovedne modele.

1.3 Motivacija

Obdelava podatkov je trenutno zelo popularno in hitro se razvijajoče področje v tehnološkem svetu. Skoraj na vsaki spletni strani lahko opazite opozorila oz. obvestila s strani skrbnikov, da morate soglašati z uporabo piškotkov med uporabo portala, saj z njihovo pomočjo izboljšujejo uporabniško izkušnjo med obiskom spletne strani. Z obdelavo podatkov se srečujemo na vsakem koraku, saj je slednja pomemben del uporabe spleta.

Pridobivanje spletnih podatkov (ang. *web scraping*) je pomemben del obdelave podatkov. Na spletu vsak dan objavimo veliko število novih podatkovnih virov. Zanesljivost vira je tista, ki določa uporabnost podatkov in smotrnost nadaljnje obdelave le-teh.

1.4 Cilji

Portal OPSI je relativno nov portal in še vedno pridobiva na prepoznavnosti. Zato je tudi obisk strani manjši. Kljub temu so nekatere podatkovne zbirke sorazmerno popularne, saj imajo že več kot 2000 uporabnikov oz. ogledov, kar je velik uspeh za eno leto star portal.

Z namenom, da bi dostopnost OPSI približali še več uporabnikom, smo razvili orodje za dostop do podatkov portala OPSI, ki omogoča uporabo podatkov v programu Orange. Izdelani vmesnik omogoča hitro in enostavno pridobivanje podatkov s portala v formatu, ki je primeren za uvoz v program Orange in nadaljnjo obdelavo.

Poglavje 2

Pregled področja in metode

Glavno tematsko področje diplomske naloge so orodja in pristopi za pridobivanje spletnih podatkov. Prikazanih je le nekaj najbolj uporabljenih tovrstnih orodij. Izpustili smo orodja, ki niso podprta v programskem jeziku Python ali pa ne ponujajo podatkov v želenih formatih.

2.1 Obstoječi portali odprtih podatkov

Portalov za dostop do odprtih podatkov je mnogo. S tematiko diplome so najbolj povezani trije portali: Open Data Slovenija, Evropski podatkovni portal in Portal odprtih podatkov EU.

2.1.1 Open Data Slovenija

Portal Open Data Slovenija [8] je zaživel leta 2012 kot del izvedbe ideje, da se širši javnosti omogoči dostop do določenega sklopa podatkov za skupno dobro. Na strani najdemo nekaj projektov, ki so rezultat uporabe podatkov iz strani z odprtimi podatki, npr., Statistični urad Slovenije [9].

2.1.2 Evropski podatkovni portal

Evropski podatkovni portal ponuja metapodatke informacij organov javnega sektorja z namenom, da se izboljša dostopnost ter vrednost odprtih podatkov. Cilj portala je promocija koristi ponovne uporabe podatkov [10].

2.1.3 Portal odprtih podatkov EU

Portal deluje od leta 2012. Omogoča ponovno uporabo podatkov in tako spodbuja gospodarski razvoj ter transparentnost v institucijah EU. Na portalu je možno iskati po katalogu podatkov ter aplikacij, ki uporabljajo podatke. Dostop do podatkov je možen preko API REST. Ponuja tudi vmesnik za poizvedovanje končne točke SPARQL. Večina podatkov je brezplačna ob navedbi vira, pri nekaterih pa veljajo za uporabo posebni pogoji [11].

2.2 Obstoječe rešitve za pridobivanje podatkov iz spleta

Za pridobivanje podatkov iz spleta obstaja vrsta orodij. Opisanih je nekaj najbolj uveljavljenih orodij.

2.2.1 Scrapy

Scrapy je aplikacijsko programsko ogrodje, ki obišče spletne strani in iz njih zgradi indeks, po katerem se lahko pregleduje podatke. S pomočjo indeksa se iz strani izlušči podatke, ki se jih lahko kasneje uporabi pri podatkovnem rudarjenju, obdelavi podatkov ali pri njihovem shranjevanju. Scrapy deluje s pomočjo pajkov, ki jih definira uporabnik in s katerimi se izlušči podatke iz spletnih strani.

2.2.2 Import.io

S spletno platformo Import.io lahko brez znanja programskih jezikov izluščimo podatke iz spletnih strani. Omogoča pretvorbo nestrukturiranih podatkov v strukturiran format, pripravljen za nadaljnjo obdelavo. Uporabnik z zaporedji klikov označi na spletni strani podatke, ki jih želi shraniti. Izbira lahko med zelo omejenim naborom formatov, v katerih je podatke možno izvoziti.

2.2.3 Web Scraper

Web Scraper je zelo podobna rešitev kot Import.io. Spletna platforma dodatno ponuja še oblačno storitev pridobivanja spletnih podatkov. Ima manjši nabor formatov izvoza podatkov kot ga ponuja Import.io.

2.2.4 Apify

Apify je spletna aplikacija za pridobivanje podatkov iz spletnih strani. Pridobivanje je možno avtomatizirati s pomočjo programa cron. Za vsako stran lahko s pomočjo orodja naredimo API. Tudi Apify ponuja oblačno rešitev pridobivanja podatkov.

2.3 Uporabljene metode pri izdelavi vmesnika

Za izdelavo vmesnika smo uporabili znane programske jezike in pripadajoče knjižnice.

2.3.1 Python

Za razvoj vmesnika smo uporabili programski jezik Python, saj ponuja širok nabora knjižnic. Za dodatno podporo pri kodiranju znakov, smo izbrali Python verzije 3. Uporabili smo knjižnice za Python:

- **Flask** je mikro programsko ogrodje za razvoj spletnih aplikacij za programski jezik Python. Ogrodje je zelo enostavno in omogoča hiter

razvoj in postavitev aplikacij.

- **requests** je namenjena pošiljanju zahtevkov HTTP v jeziku Python. Z njimi lahko prilagajamo zaglavja, parametre, vsebino ter ostale atribute zahtevkov HTTP.
- **BeautifulSoup** omogoča, da iz dokumentov HTML in XML izluščimo podatke. Z njeno pomočjo lahko pregledujemo, iščemo in spreminjamo razčlenitveno drevo dokumenta.
- **xlrd** omogoča branje tabel in podatkov shranjenih v datotekah Excel.
- **PyPDF2** se uporablja za delo z dokumenti PDF.
- **python-docx** knjižnica omogoča ustvarjanje in urejanje dokumentov Microsoft Word.
- Podatkovna zbirka **SQLite3** je najbolj uporabljeno orodje za upravljanje z relacijskimi podatkovnimi bazami. Izbrali smo jo zaradi podpore v programskem jeziku Python.

2.3.2 Brskalnik DB Browser for SQLite

Program DB Browser for SQLite je odprtokodna rešitev za vizualizacijo, shranjevanje, oblikovanje in urejanje datotek podatkovnih zbirk kompatibilnih s podatkovnimi bazami SQLite.

2.3.3 JavaScript

Ogrodje **Bootstrap 4** je eno izmed najbolj popularnih ogrodij za razvoj odzivnih ter mobilnih spletnih aplikacij. Ogrodje vključuje tehnologije HTML, CSS in JavaScript.

Knjižnica **jsTree** je napisana v Javascript in omogoča prikazovanje arhitekture podatkov na spletni strani, v obliki datotečnega sistema. Knjižnica **CanvasJS** je napisana v JavaScript in omogoča prikazovanje podatkov na platnu v obliki grafov.

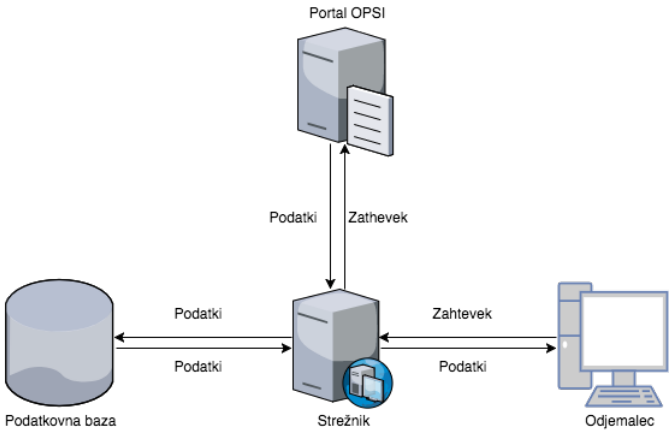
2.3.4 Git

Git je sistem za nadzor nad spremembami datotek in njihovo usklajevanje med skupinami ljudi. **GitHub** je spletno orientirano orodje Git. Ponuja tudi porazdeljeno upravljanje nad izvorno kodo v git.

Poglavje 3

Razvoj

Slika 3.1 prikazuje osnovno delovanje razvitega aplikacijskega strežnika. Na začetku je potrebno pridobiti podatke iz portala OPSI. Aplikacijski strežnik pridobiva podatke iz portala OPSI preko zahtevkov HTTP. Pridobljene podatke strežnik obdela ter jih shrani v bazo. Odjemalec-uporabnik nato zahteva podatke od aplikacijskega strežnika. Podatki, ki so pripravljeni na nadaljnjo uporabo v podatkovni bazi, so poslani do odjemalca.



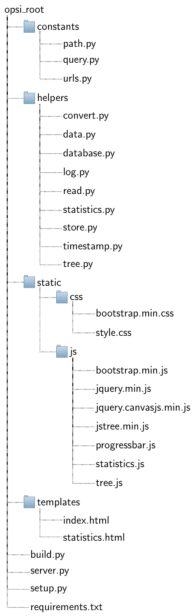
Slika 3.1: Arhitektura aplikacije.

3.1 Arhitektura vmesnika

Vmesnik lahko razdelimo na štiri različna področja:

- odjemalec,
- strežnik,
- podatkovna baza,
- portal OPSI.

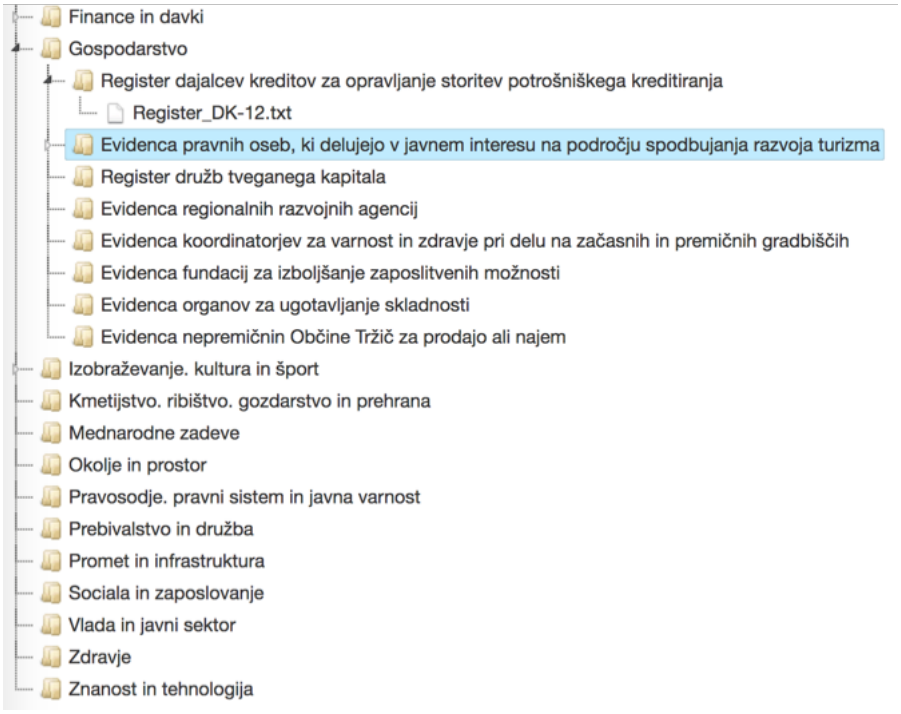
V nadaljevanju si bomo podrobneje pogledali sestavo in vlogo, ki jo opravlja vsako izmed naštetih komponent. Na spodnji sliki lahko vidimo datotečno strukturo projekta.



Slika 3.2: Datotečna struktura vmesnika.

3.1.1 Odjemalec vmesnika

V brskalnem oknu pri odjemalcu se na naslovu "https://naslov:/vrata/" prikaže osnovna drevesna struktura portala OPSI. Prikazana so področja



Slika 3.3: Prikaz osnovne drevesne strukture v vmesniku.

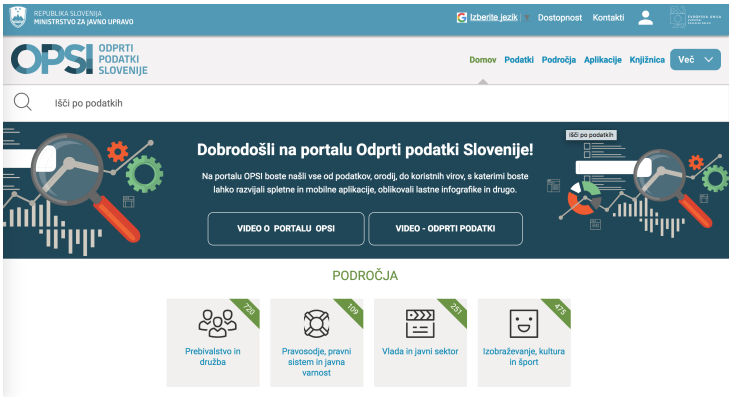
za katera so na voljo pretvorjeni podatki. Na sliki 3.3 je prikazan vmesnik, ki omogoča sprehajanje po datotekah in navideznih mapah, kot so tudi prikazani v portalu OPSI. S klikom na eno izmed datotek na vmesniku, se odpre pogovorno okno, ki nam ponudi prenos izbrane datoteke, ki je že vnaprej pripravljena za nadaljnjo obdelavo in uvoz v program Orange. Na naslovu "https://naslov:/vrata/statistics/" je prikazana statistika vseh podatkov pridobljenih iz portala OPSI. Aplikacija poroča statistiko o tipih formatov podatkov, številu napak med njihovim pridobivanjem, številu nepodprtih oz. nepoznanih formatov, deležu podatkov, ki jih je možno uvoziti v program Orange in pa številu tipov formatov za vsako področje.

3.1.2 Strežnik Flask

Strežnik Flask posluša odjemalca na vnaprej določenih končnih točkah API REST. Strežnik odjemalcu servira potrebne podatke, vse poglede, informacije o posameznih datotekah, strukturo portala ter potrebno statistiko. Prav tako ob pridobivanju podatkov opravlja pretvorbo podatkov. Pred shranjevanjem v bazo preverja, ali že obstajajo in če so zadnje verzije. Natančno kopijo podatkov iz portala shranjuje v začasni shrambo. Uporabnik tako ima vpogled tudi v izvirne, nespremenjene podatke. Strežnik poskrbi za beleženje napak, ki so nastale med posameznimi fazami delovanja.

3.1.3 Portal OPSI

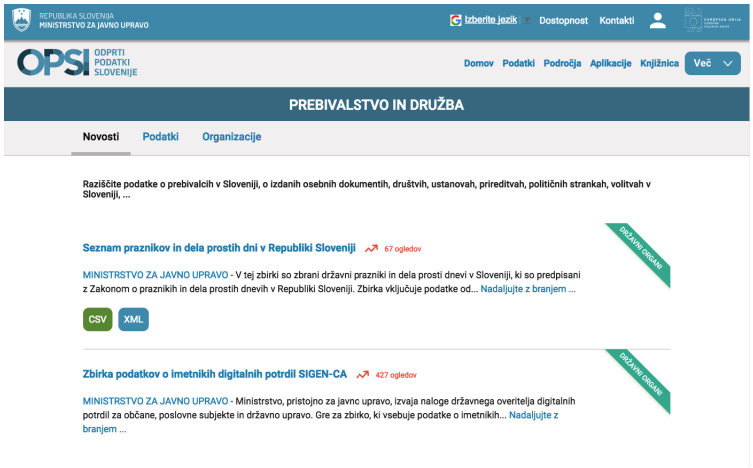
Vsi podatki, ki smo jih pripravili za nadaljnjo uporabo, izvirajo iz portala OPSI. Portal ponuja prijazen in intuitiven vmesnik, ki s svojo enostavnostjo poskrbi za dobro interakcijo uporabnika s portalom.



Slika 3.4: Uvodna stran portala OPSI.

Uvodna stran je prikazana na sliki 3.4. Izpisanih je nekaj osnovnih informacij o portalu (povezavi na predstavitvena videa). Stran omogoča iskanje po podatkovnih zbirkah s pomočjo iskalnega orodja. Pod iskalnikom se razkrijejo vsa področja podatkovnih zbirk. V zelenem pasu se izpiše število področju pripadajočih podatkovnih zbirk. Na dnu uvodne strani so prikazane

novice o portalu. Na uvodni strani je prikazanih nekaj zavihkov v navigacijski vrstici: Domov - povezava na uvodno stran, Podatki - povezava na vse podatkovne zbirke, Področja - podatkovne zbirke grupirane po področjih, Aplikacije - vse aplikacije, ki uporabljajo podatke iz portala OPSI, Knjižnica - vse novice o portalu ter trije zavihki z navodili za uporabo ter posredovanje podatkov in več informacij o portalu. Ob kliku na eno področje, v tem primeru na Prebivalstvo in družbo - 3.5, se izpiše ime vsake zbirke, tip podatkov, ki jih zbirka vsebuje in število ogledov.



Slika 3.5: Primer podatkovnih zbirk znotraj področja.

Kot je razvidno iz slike 3.6 lahko znotraj vsake zbirke ogledamo še metapodatke o sami zbirki: kratek opis, avtor, skrbnik, datum objave, številke revizije, povezave na podatke ter mnoge druge. S klikom na povezavo “Metapodatki JSON” portal vrne še vse podatke v obliki JSON.

Slika 3.6: Primer opisa vira datoteke.

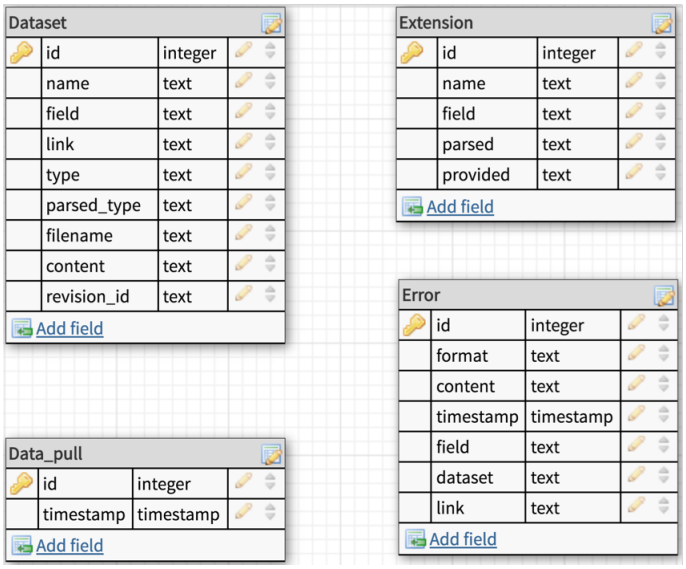
3.1.4 Podatkovna baza

Model baze je prikazan na sliki 3.7. Strežnik zapisuje v podatkovno bazo vsebino datotek in njihove metapodatke. Vsebina datotek je pred shranjevanjem obdelana v obliko za uvoz v program Orange. Baza je sestavljena iz štirih neodvisnih tabel:

- Dataset - vsebina datotek in njihovi metapodatki,
- Extension - podatki o neznanih in nepodprtih formatih datotek,
- Data_pull - časovni žig začetka vsakega pridobivanja podatkov,
- Error - podatki o napakah med pridobivanjem podatkov.

Ob pridobivanju podatkov se najprej preveri, če datoteka v bazi že obstaja. Vmesnik pogleda tabelo Dataset, kamor so zapisani metapodatki in vsebina vsake datoteke podatkovne zbirke. V primeru, da datoteka v bazi že obstaja, se preveri ali je bila popravljena ali dopolnjena, ter se posledično posodobi. V nasprotnem primeru se datoteko obdela in nato shrani v tabelo Dataset. Če pri shranjevanju pride do napake, se napako zabeleži v tabelo

Error in se jo prikaže na strani statistike o vmesniku. Ob vsakem pridobivanju podatkov se zabeležita čas in datum v tabeli Data_pull. Če vmesnik med pridobivanjem naleti na neznano ali nepodprto datoteko, jo zabeleži v tabelo Extension. Tudi ti podatki so izpisani v prikazu statistike.



Slika 3.7: Model podatkovne baze.

3.2 Postavitev okolja

Pred začetkom razvoja je bilo potrebno najprej postaviti funkcionalno okolje ter namestiti vse knjižnice, ki so potrebne za delovanje vmesnika. Namestili smo programski jezik Python verzije 3.6.3, ki je osnova za razvoj vmesnika. Namestili smo tudi sistem za upravljanje paketov pip, s katerim smo kasneje namestil še ostale potrebne knjižnice:

- requests
- BeautifulSoup
- xlrd

- PyPDF2
- python-docx

Na koncu je bilo potrebno namestiti še program "DB Browser for SQLite" za pregled podatkovne baze.

3.3 Razvoj vmesnika

Vmesnik je sestavljen iz čelne ter zaledne aplikacije. Poleg opisa razvoja so omenjene tudi uporabljene tehnologije.

3.3.1 Čelni del aplikacije

Grafični uporabniški vmesnik smo razvili z uporabo tehnologije HTML5, CSS in Javascript v kombinaciji z ogrodjem Bootstrap 4 za kreiranje odzivnih spletnih aplikacij. Zahtevki iz odjemalčeve strani so definirani s knjižnico jQuery in sicer z Ajax-ovimi zahtevki HTTP. Vsa področja, podatkovne zbirke ter pripadajoče datoteke iz portala OPSI so prikazane na uporabniku prijazen in že poznan način - v obliki datotečnega sistema. Za grajenje datotečnega sistema oz. drevesa smo uporabili knjižnico jsTree, ki se je izkazala kot zelo primerna in dobro služi svojemu namenu. Prikaz statistike nad podatki je implementiran s knjižnico CanvasJS, s katero se enostavno izdela tortne ter stolpične diagrame. Koda se nahaja v direktoriju "static," ki si naprej veji na poddirektorije "js" in "css," v katerih se nahajajo skripte, ki uporabljajo vse potrebne knjižnice za normalno delovanje in prikaz vmesnika.

3.3.2 Zaledni del aplikacije

Izvorna koda strežniške logike je v celoti napisana v Flask - mikro ogrodje za razvoj spletnih aplikacij za Python. Izvorna koda je razdeljena na več paketov oz. angleško "package directory." Datoteke s podobnimi funkcionalnostmi so

grupirane v istih direktorijih. Ogrodje Flask zahteva določeno strukturo projekta. Vse datoteke HTML datoteke je potrebno vstaviti v direktorij "templates," saj naj bi te predstavljale vnaprej pripravljene predloge za izgled spletne aplikacije in jih je potrebno grupirati. Naslednji obvezni direktorij je "static," kamor vstavimo vse ostale potrebne datoteke, kot so slike, JavaScript ali CSS koda ter njihove knjižnice, ki jih poljubno vključujemo v projekt. V korenu projekta je datoteka s strežniško logiko "server.py," ki zažene aplikacijo ter uporablja dodatna modula "constants" in "helpers." Skripta "server.py" prav tako komunicira z zaledjem aplikacije preko vmesnika API REST. V modulu "constants" najdemo razne konstante ter povezave, spletne ali fizične. Modul "helpers" vsebuje vse dodatne potrebne funkcionalnosti za upravljanje baze, beleženje napak, beleženje statistike, pretvarjanje podatkovnih tipov in še več. V korenu projekta se nahaja skripta "build.py," ki je zadolžena za sprehajanje po portalu OPSI ter prenos podatkov in njihovo pretvorbo v obliko, ki je kompatibilna s programom Orange. Vsi podatki se shranijo v podatkovno bazo in hkrati tudi fizično v začasno shrambo. Čelni del oz. "server.py" z zaledjem aplikacije komunicira preko REST, z zahtevki v formatu json. Na korenskem nivoju skripta "setup.py" poskrbi, da se pred zagonom skripte "build.py" namestijo vse potrebne knjižnice za delovanje aplikacije.

3.4 Algoritem določanja tipa atributov

Da bi lahko vsebino datotek uspešno uvozili v program Orange, jih je potrebno sprva obdelati in pripraviti v ustrezno obliko. Za datoteke, ki so formata csv, xls ali xlsx je potrebno določiti tip podatkov v posameznem stolpcu. V vrstico pod imenom vsakega stolpca je potrebno vpisati eno izmed štirih možnih tipov: "t" za časovno vrsto, "s" za niz znakov, "d" za diskretne in "c" za zvezne spremenljivke.

Algoritem za pretvorbo datotek formatov csv, xls in xlsx najprej prebere vsebino datoteke v dvodimenzionalno tabelo. Tabelo nato transponira in

tako vrstice zamenja s stolpci. Algoritem nato pregleda posamezno vrstico, ki vsebuje podatke za vsak stolpec in preverja po naslednjemu zaporedju: če je možno vsa polja v vrstici pretvoriti v časovno vrsto, potem je vrstica časovna vrsta, dodati je potrebno znak "t". Če prejšnji pogoj ni bil izpolnjen, preverimo ali vsa polja vsebujejo diskretne podatke in v pozitivnem primeru dodamo oznako "d". Sicer se premaknemo na naslednji pogoj. Če se v vseh poljih zvezni podatki, označimo vrstico z znakom "c". V nasprotnem primeru podatke identificiramo kot nize znakov. Preden algoritem vrne tabelo, jo še enkrat transponira v pravo obliko za Orange.

3.5 Testiranje in težave

Med razvojem vmesnika smo naleteli na določene težave. Razvili smo tudi nekaj testov, s katerimi preverjamo ustreznost delovanja vmesnika.

3.5.1 Kompatibilnost

Vmesnik smo testirali na različnih tipih brskalnikov. Vmesnik preverjeno deluje na brskalnikih Google Chrome, Mozilla Firefox, Safari ter Internet Explorer 11. Preverili smo prikazovanje grafov v različnih brskalnikih. Opaziti ni bilo občutnih razlik ali težav pri izrisu statistik in uporabi vmesnika, ki se povsod obnaša enako in deluje normalno.

3.5.2 Testiranje pravilnosti podatkov

Da bi lahko potrdil pravilnost delovanja vmesnika, smo testirati tudi uvoz pripravljenih podatkov v program Orange. Uspešnost uvoza vseh tipov formatov datotek smo preverili ročno tako, da smo izbrali 10% vseh možnih datotek kot testno množico in vse uspešno uvozili v program Orange.

3.5.3 Težave med izdelavo vmesnika

Med samo izdelavo vmesnika nismo naleteli na veliko težav, so pa bile dokaj kompleksne. Težavo na portalu OPSI predstavlja že njegov vmesnik API za metapodatke. Podatke vrača v formatu json. Eden izmed mnogih atributov je tudi atribut update_date, ki naj bi predstavljal datum zadnje posodobitve podatkov. Vrednost atributa se spreminja vsak dan ob približno peti uri zjutraj, čeprav se podatki ne spremenijo. Sprememba podatkov je zapisana posredno v atributu revision_id, ki se je kasneje izkazal kot ključni element za spremljanje ažurnosti podatkov v podatkovni bazi preden se le-ti posodobijo. Naslednji problem so napačno klasificirani formati datotek. Mnogo datotek je na portalu OPSI klasificiranih kot eden izmed formatov, ki jih vmesnik podpira. Kasneje se izkaže, da so podatki ali arhivirani ali pa popolnoma drugega formata, ali pa so celo kombinacija dveh formatov. Nemogoče je spisati rešitev, ki bo za vse izmed teh možnih napak oz. težav našla optimalno rešitev in datoteko pravilno klasificirala, zato smo vse take primere uvrstil kot napake med pridobivanjem podatkov.

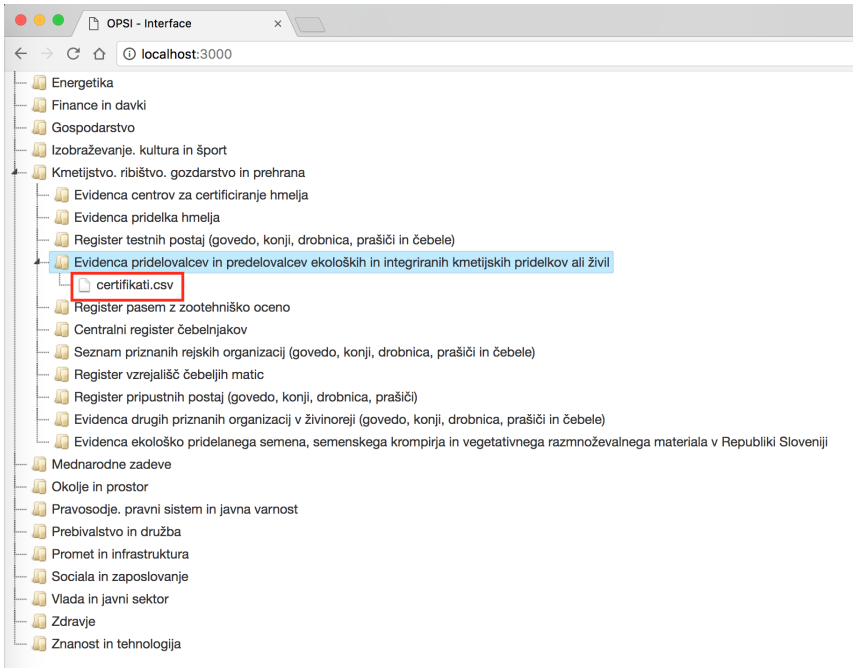
Poglavje 4

Rezultati

Izvorna kodo razvoja vmesnika je dostopna na naslovu `https://github.com/saso285/OPSI`, kjer si lahko preberemo tudi kratka navodila za namestitev vmesnika in njegovega upravljanja. Vmesnik naredi obhod po celotnem portalu OPSI in hkrati med pridobivanjem podatke še obdela, shrani v bazo in lokalno na disk. Zbirska obsega 1 GB podatkov. Proces pridobivanja podatkov z internetno povezavo 100/10 Mbps traja približno pol ure. Podatki, ki pa so že bili obdelani in shranjeni v bazo so dostopni za prenos iz strežnika še preden se postopek pridobivanja zaključi.

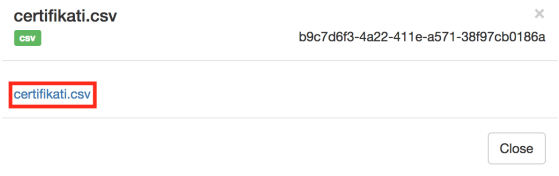
4.1 Primer uvoza podatkov iz vmesnika v program Orange

Delovanje procesa uvoza podatkov iz vmesnika v program Orange je prikazano na primeru, ki vključuje uvoz datoteke formata csv. Če v brskalniku odpremo vmesnik, se bo prikazala struktura podobna tisti na sliki 4.1. Na sliki lahko opazimo, da je v vmesniku izbrana pot do datoteke certifikati.csv znotraj področja Kmetijstvo, ribištvo, gozdarstvo in prehrana.



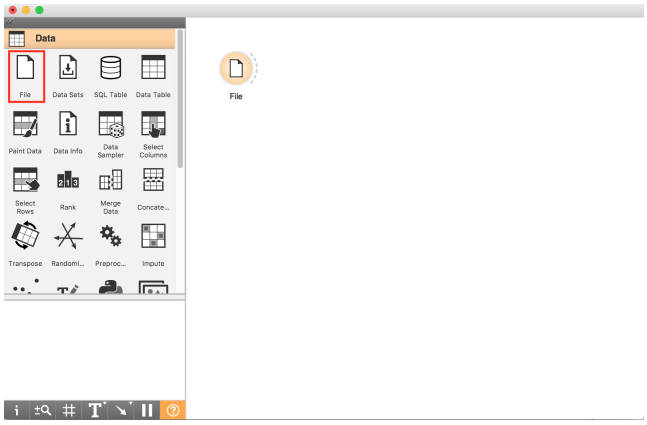
Slika 4.1: Pregled vejanja strukture vmesnika.

S klikom na datoteko certifikati.csv se odpre pogovorno okno kot je razvidno na sliki 4.2, kjer si lahko uporabnik prenese želeno datoteko na svoj računalnik.



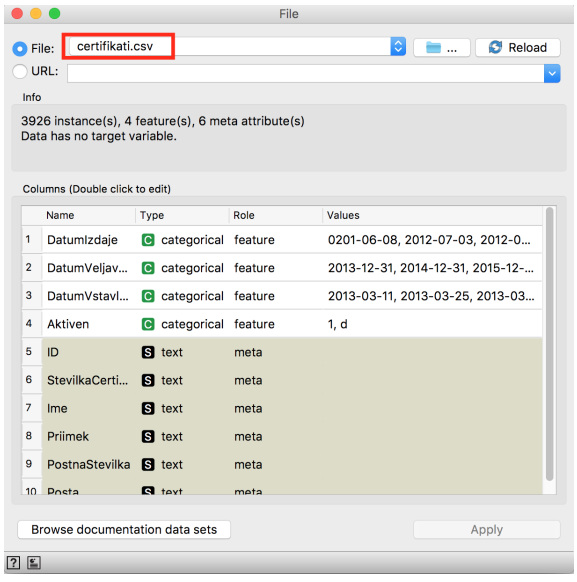
Slika 4.2: Pogovorno okno za prenos datoteke.

Po prenosu datoteke, je potrebno zagnati program Orange. Ko kreiramo nov projekt, kot na sliki 4.3, s klikom na ikono File v stranski orodni vrstici ustvarimo na delovni površini nov gradnik.



Slika 4.3: Glavno okno programa Orange.

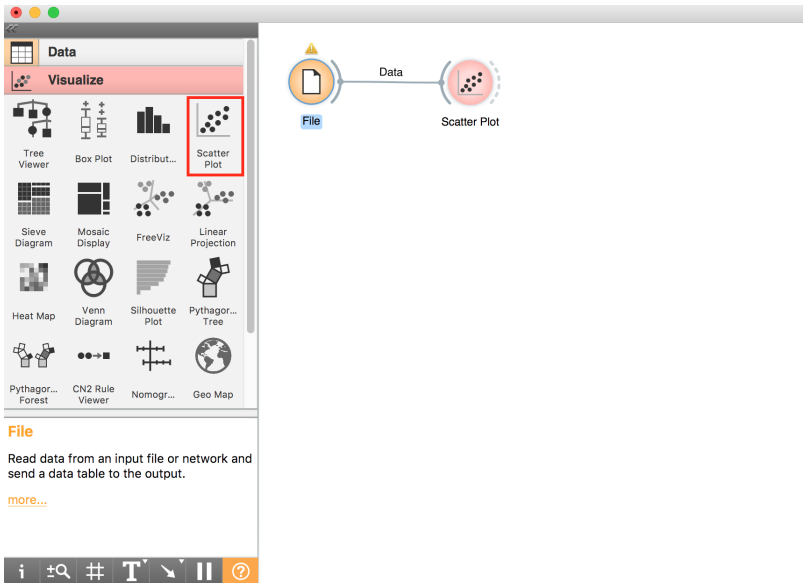
Z dvojnim klikom na novo ustvarjeno komponento File na delovni površini, odpremo novo okno, kjer lahko izberemo datoteko za uvoz v program orange, kot je prikazano na sliki 4.4.



Slika 4.4: Uvoz nove datoteke v program Orange.

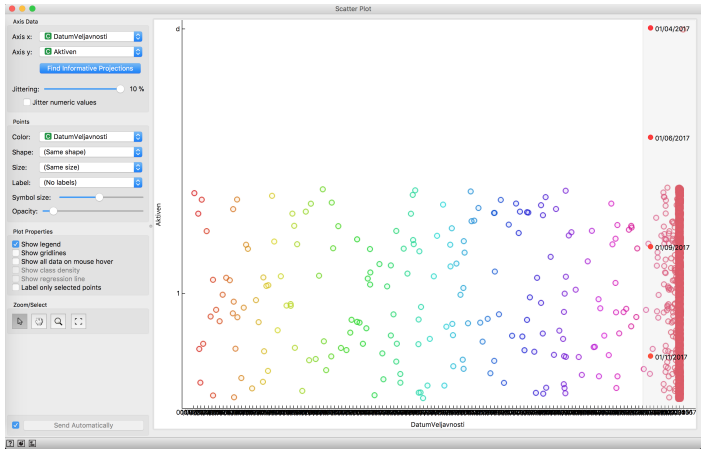
Uporabnik lahko po uspešnem uvozu podatov v komponento File začne z nadaljnjo obdelavo in vizualizacijo podatkov. Zato dodajmo še dve doda-

tni komponenti za prikaz podatkov (zbirka Visualize). Prikaz Distributions pokaže porazdelitve vrednosti podatkovnih značilnosti v obliki grafa. S prikazom Scatter Plot lahko vizualiziramo podatke tudi na koordinatnem sistemu. Komponente nato medsebojno povežemo in rezultat prejšnjih akcij si lahko ogledamo na sliki 4.5.



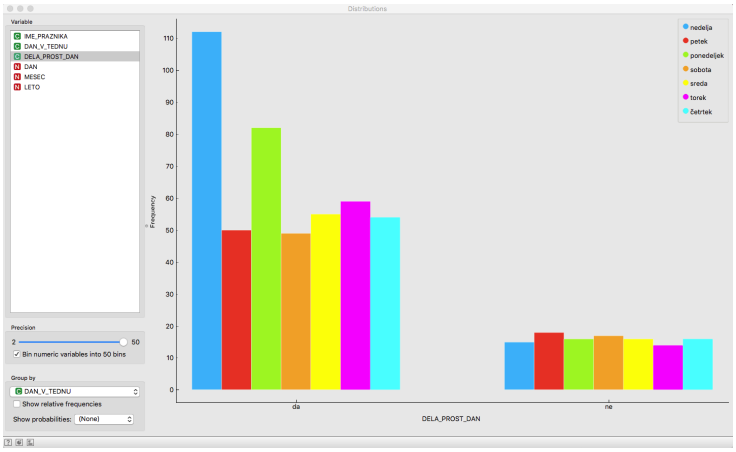
Slika 4.5: Povezava komponente File s komponentama Distributions in Scatter Plot.

Tako smo s formalnostmi zaključili in se lahko osredotočimo na podatkovno vizualizacijo ter nadaljnje delo s podatki, kar je razvidno tudi na sliki 4.6.



Slika 4.6: Prikaz rezultata komponente Scatter Plot.

Drug primer uvoza na isti način kot v zgornjem primeru predstavlja datoteka “seznampraznikovindelaprostihdni20002030.csv”. Podatkovna zbirka Seznam praznikov in dela prostih dni v Republiki Sloveniji spada v področje Prebivalstvo in družba in vsebuje seznam praznikov ter dela prostih dni vse od leta 2000 do 2030. Na sliki 4.7 si lahko ogledamo histogram, ki ima na abcisni osi vrednost ali je nek dan dela prost ali ne, na ordinatni osi pa je frekvenca dni v tednu, ki ustrezajo posamezni kategoriji (dela prost, delovni dan).

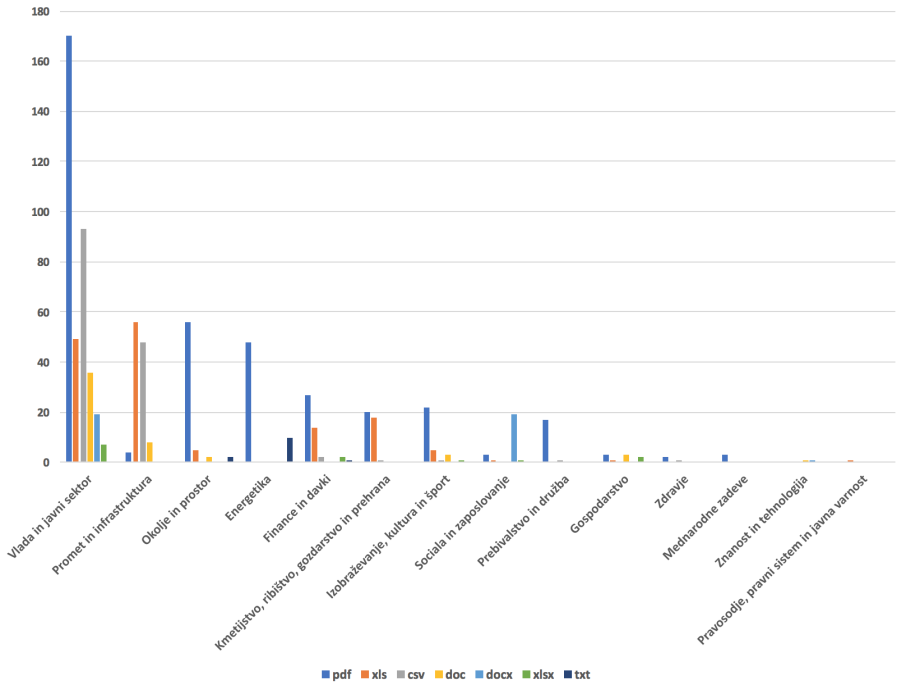


Slika 4.7: Histogram porazdelitve pogostosti praznika na določen dan v tednu.

4.2 Uspešnost detekcije in pretvorbe podatkov

Razviti vmesnik omogoča dostop do relativno majhnega odstotka razpoložljivih virov datotek. Uspešno jih pretvori le 4% (788 od 19565). Glavni razlog za nizek procent uporabniku razpoložljivih virov datotek je veliko število datotek formata html, ki se ne shranjujejo v bazo, saj vmesnik ne podpira datotek tega tipa. Prav tako se število dostopnih virov datotek zmanjša za nekoliko procentov zaradi napak, ki nastanejo pri samem pridobivanju podatkov.

Iz zgornjega odstavka je razvidno, da niso vsi formati virov datotek uporabni. Najbolj nazoren primer je seveda format html, pri katerem ne moremo vedeti ali nam bo vir ponudil povezave na dodatne datoteke, ali je na drugi tabela s podatki, besedilo ali pa kar celo iskalnik po spletni strani. Tukaj se pojavi dilema, kaj obravnavati kot končne podatke vira. Drug zanimiv podatek je tudi to, da je med vsemi dostopnimi datotekami največ tistih tipa pdf (383 datotek oz. slabih 49%). Izstopata še format xls s 150 datotekami in format csv s 147 datotekami. Preostali formati imajo občutno manjše število



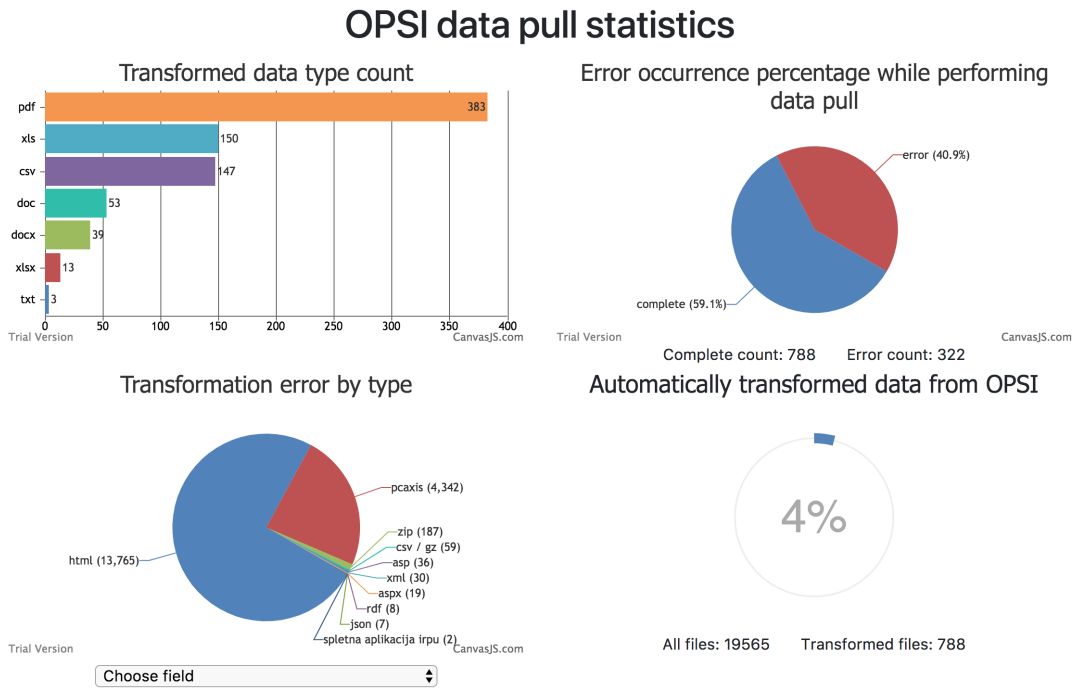
Slika 4.8: Histogram formatov virov datotek glede na posamezno področje.

virov datotek.

Na sliki histograma 4.8 lahko opazimo, da največ podatkov izhaja iz področja Vlada in javni sektor. Slednje vsebuje 375 virov datotek, kar predstavlja približno 48% vseh uporabniku razpoložljivih podatkov. Sledita še področji Promet in infrastruktura s 116 viri datotek ter Okolje in prostor z občutno manjšim številom in sicer 65 virov datotek.

Med pridobivanjem datotek iz portala OPSI se pri približno 15% datotek pojavijo napake. Vzroki za napake so različni: napake v kodiranju datotek, težave pri prenosu datotek, itd.

Največ virov datotek, ki so pripravljeni za prenos iz vmesnika, je iz področja Vlada in javni sektor, in sicer 281. Iz preostalih področij ostaneta le še dve, ki malo izstopata: Promet in infrastruktura z 80 datotekami ter Okolje in prostor z 62 datotekami.



Slika 4.9: Zavihek statistika podatkov za vmesnik.

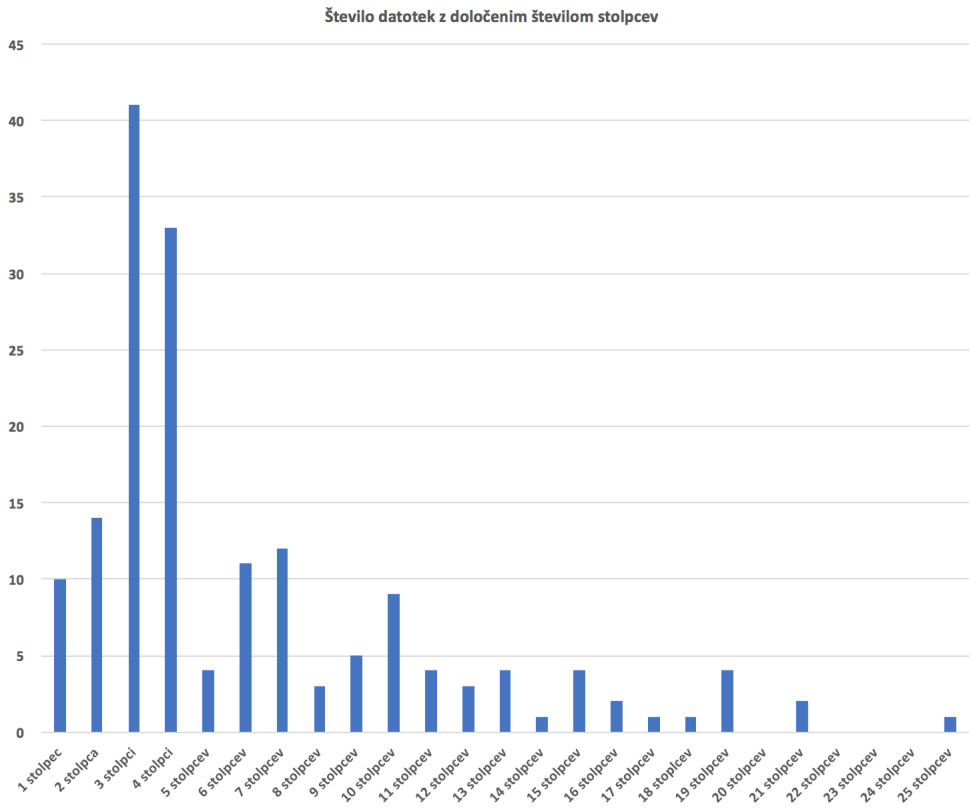
Vseh datotek s preglednicami je 310. Vštete so vse datoteke formata CSV, XLS in XLSX. Približno 15% teh datotek ni pravilno strukturiranih, kar pomeni, da se iz njih ne da razbrati tabel ali pa so sestavljene v kombinaciji s slikami in od njih nimamo koristi. Če odštejemo od vseh datotek s preglednicamo vse neustrezne datoteke, dobimo približno 260 uporabnih datotek.

V programu Orange pri uvozu datotek s preglednicami obstaja več tipov spremenljivk za identificiranje stolpcev in sicer:

- continuous - zvezne spremenljivke (neskončno število možnih vrednosti)
- discrete - diskretne spremenljivke (končno število možnih vrednosti)
- string - niz znakov
- time - časovni podatki

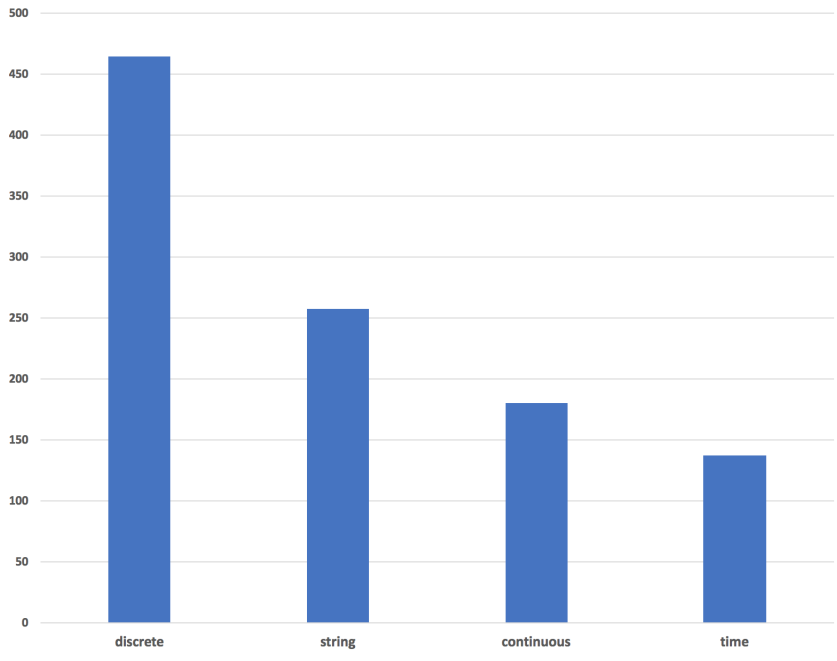
Slednje lahko identificiramo le za vire datotek, ki so strukturirane kot preglednice oz. se jih da uvesti v urejevalnik preglednic.

Število vseh stolpcev v 260 datotekah s preglednicami je 1038. Slika 4.10 prikazuje razporeditev datotek z določenim številom stolpcev. Naključno izbrana datoteka s preglednico vsebuje tri stolpce in sicer eden izmed stolpcev je tipa discrete, druga dva pa sta izmed preostalih treh tipov. Obstajajo seveda tudi izjeme, kjer datoteka s preglednicami vsebuje več različnih stolpcev, kot je lahko razvidno iz zgornjih dveh primerov uvoza podatkov in njim pripadajočih slik.

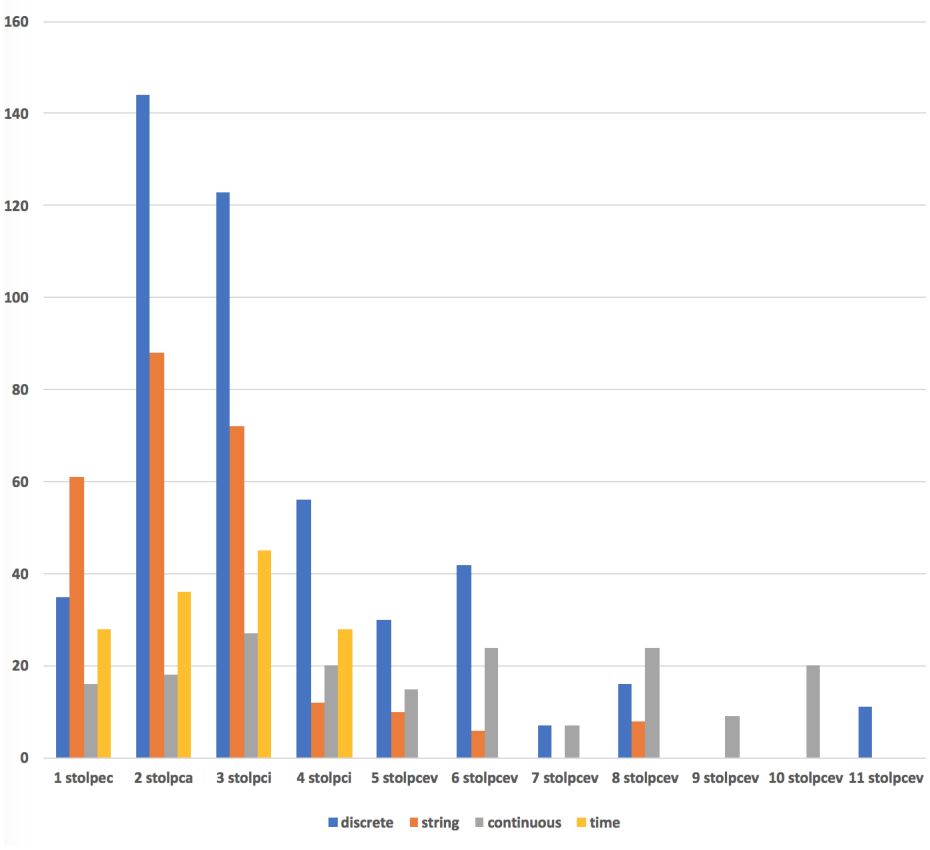


Slika 4.10: Histogram porazdelitve datotek, ki vsebujejo določeno število stolpcev.

Na sliki 4.11 je prikazano tudi število posameznih spremenljivk v vseh virih datotek s preglednicami. Največ stolpcev je tipa discrete in sicer 464, število ponovitev ostalih treh tipov spremenljivk je dokaj nižje in variira nekje med 257 ter 137.



Slika 4.11: Histogram ponovitev tipov spremenljivk v datotekah s preglednicami.



Slika 4.12: Histogram ponovitev števila različnih tipov stolpcev v datotekah s preglednicami.

Poglavje 5

Sklepne ugotovitve

Portal OPSI je neglede na nizek delež podatkov, ki je možno uvoziti v program Orange, še vedno odličen vir podatkov, saj imajo vsi podatki zanesljiv izvor. Podprtost formatov podatkov se da dodatno razširiti in tako povečati število datotek pripravljenih za nadaljnje delo. Potrebno je tudi omeniti, da so datoteke, ki so podprte s strani našega vmesnika dobro formulirane in so ustrezne za delo tako za laike kot tudi za strokovnjake.

Mnogo formatov datotek, ki jih ponuja portal OPSI, je neuporabnih za uvoz in nadaljnjo obdelavo v programu Orange. To so formati: pcaxis, json, asp/x ter html. Največji problem predstavlja format html, saj so viri slednjega tipa le povezave do besedil, tabel, zemljevidov ali pa celo do drugih strani, ki vsebujejo povezave za prenos končnih datotek. Tako je do končnih podatkov zelo težko priti, saj se soočamo z variabilnimi globinami na katerih se slednji nahajajo. Prav tako pa naletimo na težave pri sami detekciji končnih podatkov v virih datotek formata HTML.

Nadaljnji razvoj vmesnika bo obsegal predvsem dodajanje nekaterih izmed nepodprtih formatov datotek, ki imajo potencial za pretvorbo v format, ki je kompatibilen s programom Orange. Na ta način bomo razširili raznolikost in povečali količino podatkov, ki bodo na voljo za nadaljnjo obdelavo in vizualizacijo.

Poleg dodajanja novih formatov bi bilo smiselno dodati statistike o podat-

kih. Osnovna statistika opisuje le določene lastnosti razpoložljivih podatkov. Zato je smiselno vložiti še nekaj dodatnega truda in dela za analizo pridobljenih podatkov in povečanje uporabnosti podatkov. Tako bomo tudi motivirali morebitne nove ali in obstoječe uporabnike za nadaljnjo uporabo podatkov iz portala OPSI.

Trenuten uvoz podatkov iz vmesnika v program Orange poteka preko datotek, ki jih uporabnik pridobi z razvitim vmesnikom. S preprostim gradnikom za program Orange, ki bi omogočal neposreden uvoz podatkov iz OPSI v Orange, bi prihranili čas in izboljšali uporabniško izkušnjo.

Literatura

[1] J. Štebe, "Odprti podatki, načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v sloveniji," *Bilten*, vol. 57/XXXVI, pp. 21–22, 2014.

[2] P. Walk, "Odprti podatki." Dosegljivo: `https://sl.wikipedia.org/wiki/Odprti_podatki`, 2009.

[3] A. Veršič, "Odprti podatki v sloveniji (kje, kam in kako)." Dosegljivo: `http://www.geoportal.gov.si/img/novice/2903_161108_7-SLO_INSPIREdan_OdprtiPodatki-Versic.pdf`, 2016.

[4] "Odpri podatki slovenije." Dosegljivo: `https://podatki.gov.si/o-portalu`.

[5] J. Štebe, S. Bezjak, and I. Vipavc Brvar, "Priprava raziskovalnih podatkov za odprti dostop : priročnik za raziskovalce," 2015.

[6] *Odprti podatki : načrt za vzpostavitev sistema odprtega dostopa do raziskovalnih podatkov v Sloveniji*. Ljubljana: Fakulteta za družbene vede, 2013.

[7] J. Demšar, T. Curk, A. Erjavec, Č. Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, *et al.*, "Orange: data mining toolbox in python," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2349–2353, 2013.

[8] D. Kožar, "opendata.si." Dosegljivo: `https://opendata.si/`, 2017.

[9] "Statistični urad republike slovenije." Dosegljivo: http://www.stat.si/StatWeb/.

[10] "Evropska unija - odprti podatki." Dosegljivo: https://europa.eu/european-union/documents-publications/open-data_sl.

[11] "Portal odprtih podatkov eu." Dosegljivo: https://data.europa.eu/euodp/sl/about.