

UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Miha Zidar

Dostop do podatkov Svetovne banke v orodju Orange

DIPLOMSKO DELO
UNIVERZITETNI ŠTUDIJSKI PROGRAM RAČUNALNIŠTVO
IN INFORMATIKA

MENTOR: prof. dr. Blaž Zupan

Ljubljana, 2016

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

Besedilo je oblikovano z urejevalnikom besedil \LaTeX .

Namesto te strani **vstavite** original izdane teme diplomskega dela s podpisom mentorja in dekana ter žigom fakultete, ki ga diplomant dvigne v študentskem referatu, preden odda izdelek v vezavo!

IZJAVA O AVTORSTVU DIPLOMSKEGA DELA

Spodaj podpisani Miha Zidar, z vpisno številko **63060317**, sem avtor diplomskega dela z naslovom:

Dostop do podatkov Svetovne banke v orodju Orange

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom prof. dr. Blaža Zupana,
- 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. avgust 2016

Podpis avtorja:

Zahvalil bi se mentorju, prof. dr. Blaža Zupana, za pomoč in usmerjanje med izdelavo diplomskega dela. Prav tako bi se za spodbudo zahvalil svojim staršem in prijateljem.

Kazalo

Povzetek

Abstract

1	Uvod	1
1.1	Motivacija	2
1.2	Cilji in struktura diplomske naloge	2
2	Spletni viri indikatorjev drzav sveta	3
2.1	Podatki indikatorjev razvoja drzav	4
2.2	Tezave pri dostopu	8
3	Knjižnica in gradniki za Orange	9
3.1	Knjiznica simple_wbd	9
3.2	api_wrapper	11
3.3	Grafični vmesnik	12
4	Uvod	13
4.1	Spletni viri	13
5	Sklepne ugotovitve	15

Povzetek

Naslov: Dostop do podatkov Svetovne banke v orodju Orange

Avtor: Miha Zidar

Program Orange je prosto dostopno orodje za podatkovno rudarjenje, s katerim lahko za namene analiz uporabimo različne podatkovne vire. Sam program Orange vsebuje lastne testne predpripravljene podatkovne vire, podobne vire si lahko pripravi in uvozi tudi uporabnik sam, ali pa uporabi katerega od že obstoječih dodatkov za uvoz podatkov. Za namen naloge smo izdelali dodatek Orange data sets (ODS), s katerim je mogoče prebrati podatke s prosto dostopnega programskega vmesnika (API) Svetovne banke (SB). Trenutno Svetovna banka omogoča uporabo starih različnih API-jev (gospodarski indikatorji (time-series), projekti SB, finančni in klimatski podatki). Dodatek ODS je namenjen lažjemu branju in pretvorbi podatkov indikatorjev in klimatskih podatkov. S tem bo uporabnikom programa Orange omogočena enostavnejša uporaba velikega števila podatkov iz omenjenih dveh programskih vmesnikov.

Ključne besede: Podatkovno rudarjenje, programski vmesnik, svetovna banka, gospodarski indikatorji, podnebni podatki, Orange.

Abstract

Keywords:

rainbow

Poglavje 1

Uvod

Na svetovnem spletu je dosegljivih vedno več prosto dostopnih programskih vmesnikov (API, ang. application programming interface). Ti vmesniki omogočajo dostop do zelo raznolikih baz podatkov. Nekateri primeri baz so sezname stopnje ogroženosti živali po državah ¹, nasini podatki meritev in slike vesolja ², seznam knjig z ocenami in povezavami med uporabniki ³, zgodovina meteoroloških meritev, ⁴, razni indikatorji stopenj razvoja držav ⁵.

Programski vmesniki so oblikovani tako, da je omogočena raznolika uporaba podatkov iz podatkovnih baz. To pa ima tudi slabost, ki je v tem, da je podatke potrebno predhodno obdelati za vsak namen posebej. Tako bi na primer moral vsak uporabnik programa Orange podatke predhodno pretvoriti v obliko, primerno za njegovo konkretno analizo.

¹<http://apiv3.iucnredlist.org/api/v3/docs>

²<https://api.nasa.gov/>

³<https://www.goodreads.com/api>

⁴<http://climatedataapi.worldbank.org/>

⁵<http://api.worldbank.org/>

1.1 Motivacija

Povezava programskega vmesnika in orodja za analizo podatkov je pogosto prezapletena za navadnega uporabnika. Z dodatkom ODS zelimo podatke programske vmesnika Svetovne banke spraviti v obliko, primerno za nadaljno uporabo v orodju Orange. Ta dodatek bi pomagal združiti programe za obdelavo podatkov in prosto dostopnih baz podatkov. S tem dobimo enostavnejši dostop do podatkov iz prek 16.000 indikatorjev in stevilnik klimatskih meritev, s čimer bomo lažje analizirali in iskali morebitne zakonitosti med podatki. V kolikor bi imeli en sam ustrezen dodatek (add-on) za dostop do podatkov programskega vmesnika Svetovne banke, bi poenostavili posodabljanje in vzdrževanje kode v primeru sprememb programskega vmesnika za vse uporabnike istega orodja hkrati. S tem odpravimo potrebo, da bi moral vsak uporabnik sam skrbeti za uskladitvene posodobitve.

1.2 Cilji in struktura diplomske naloge

Cilj diplomske naloge je izdelati knjižnico za uporabo programskega vmesnika Svetovne banke ter izdelati dodatek za program Orange, ki s pomočjo omenjene knjižnice omogoča uporabniku dostop do podatkov SB preko grafičnega vmesnika.

V drugem poglavju diplomskega dela predstavimo spletne vire indikatorjev držav sveta. Nato bomo podrobneje opisali programski vmesnik za dostop do podatkov Svetovne banke (API SB). V četrtem poglavju sledi predstavitev knjižnice in gradnikov za Orange in nato se konkretni primeri uporabe. Na koncu bomo popisali opravljeno delo, navedli vire kode in opisali nadaljne možnosti nadgradnje dodatka.

- Prav tako se večina programov in knjižnic za dostop do baz podatkov osredotoči le na iskanje po teh bazah, ne pa tudi na pridobivanje čim večje količine podatkov.

Poglavje 2

Spletni viri indikatorjev držav sveta

Na spletu je mnogo prosto dostopnih virov oz. baz podatkov. Ti imajo programske vmesnike, ki omogočajo dostop do raznovrstnih podatkov, kot so npr. sezname stopnje ogroženosti zivali po državah ¹, Nasini podatki meritev in slike vesolja ² seznam knjig z ocenami in povezavami med uporabniki ³, zgodovina meteoroloških meritev ⁴, razni indikatorji stopnje razvoja držav ⁵. Pri nalogi smo se osredotocili na dva programska vmesnika za dostop podatkov Svetovne banke, to sta zgodovina meteoroloških meritev (4) in razni indikatorji stopnje razvoja držav (5).

Za podatkovno bazo Svetovno banko smo se odločili, ker združuje in na enovit način predstavi podatke iz več različnih virov. Podatkovni viri za indikatorje stopnje razvoja držav so:

- World Development Indicators, - Global Development Finance, - African development Indicators, - Doing Business, - Enterprise Surveys, - Millennium Development Goals, - Education Statistics, - Gender Statistics, - Health and

¹<http://apiv3.iucnredlist.org/api/v3/docs>

²<https://api.nasa.gov/>

³<https://www.goodreads.com/api>

⁴<http://climatedataapi.worldbank.org/>

⁵<http://api.worldbank.org/>

Nutrition Statistics, - IDA Results Measurement System.

Podatkovni viri za klimatske meritve so pridobljeni s svetovnih meteoroloških postaj.

Dostop do podatkov je omogočen prek vmesnika REST, ki ponuja veliko možnosti za iskanje in presejanje rezultatov.

Note: vse zahteve so omejene s količino podatkov na stran in za dostop do vseh se moramo sami sprehoditi čez vse strani.

format zahtev:

```
[  
  
  {  
    "page": 1,  
    "pages": 6,  
    "per\_page": "2000",  
    "total": 304  
  },  
  [  
    {<podatki>}  
  ]  
]
```

2.1 Podatki indikatorjev razvoja držav

Programski vmesnik indikatorjev razvoja držav Svetovne banke omogoča dostop do prek 16.000 raznih indikatorjev. Podatki indikatorjev so merjeni od leta 1960 dalje v mesečnem, cetrletnem ali letnem intervalu.

Za dostop do podatkov posameznega indikatorja, potrebujemo identifikator (id) indikatorja in kodo države ali regije.

2.1.1 Opis programskega vmesnika indikatorjev

Seznam vseh indikatorjev z imeni, opisi, in identifikatorji lahko dobimo na naslovu <http://api.worldbank.org/indicators>.

Primer podatkov indikatorja "Poverty Headcount (\$2.50 a day)" (stopnja revscine pri dohodku 2,5 dolarja na dan) v obliki JSON.

```
{
  "id": "1.0.HCount.2.5.usd",
  "name": "Poverty Headcount (\$2.50 a day)",
  "source": {
    "id": "37",
    "value": "LAC Equity Lab"
  },
  "sourceNote": "The poverty headcount index measures the proportion of
    population with daily per capita income (in 2005 PPP)
    the poverty line.",
  "sourceOrganization": "LAC Equity Lab tabulations of SEDLAC (CEDLAS and
    World Bank).",
  "topics": [
    {
      "id": "11",
      "value": "Poverty"
    }
  ]
}
```

2.1.2 Opis programskega vmesnika drzav

Seznam kod drzav in regij se nahaja na naslovu <http://api.worldbank.org/countries>.

Primer

```
{
```

```
"id": "ABW",
"iso2Code": "AW",
"name": "Aruba",
"region": {
  "id": "LCN",
  "value": "Latin America & Caribbean "
},
"adminregion": {
  "id": "",
  "value": ""
},
"incomeLevel": {
  "id": "HIC",
  "value": "High income"
},
"lendingType": {
  "id": "LNX",
  "value": "Not classified"
},
"capitalCity": "Oranjestad",
"longitude": "-70.0167",
"latitude": "12.5167"
},
```

Kode drzav, ki jih lahko uporabimo v poizvedbah, ustrezajo standardu ISO 3166-1 ⁶ alpha-2 ali ISO 3166-1 alpha-3. Poleg tega gornji seznam drzav vsebuje tudi dodatne kode za agregate, kot so: - regije - skupine drzav po visini dohodka - nekatere izjeme drzav, kot je trenutno Kosovo.

⁶http://www.iso.org/iso/country_codes.htm

2.1.3 Dostop do podatkov indikatorjev

Dostop do podatkov posameznega indikatorja je mogoc na naslovu <http://api.worldbank.org/countries> kjer je:

- country_list; z podpicjem locen seznam kod drzav, ki jih dobimo v polju "id" ali "iso2Code" v zgoraj opisanem seznamu drzav, - indicator_id; identifikacijska koda (id) indikatorja.

Izsek podatkov:

```
...
{
  "indicator": {
    "id": "SP.POP.TOTL",
    "value": "Population , total"
  },
  "country": {
    "id": "IL",
    "value": "Israel"
  },
  "value": "6289000",
  "decimal": "0",
  "date": "2000"
},
{
  "indicator": {
    "id": "SP.POP.TOTL",
    "value": "Population , total"
  },
  "country": {
    "id": "IT",
    "value": "Italy"
  },
}
```

```
"value": "56974100",  
"decimal": "0",  
"date": "2001"  
  
},  
...
```

Vmesnik indikatorjev SB omogoča poizvedbe z naslednjimi parametri:

- MRV (most rescent value) - nekaj zadnjih meritev
- frequency - pogostost vzorčenja (letno, cetrletno, mesечно)
- gapfill - manjkajoče vrednosti prejsnjih meritev
- date - datum ali obdobje
- page - stran
- per_page - število elementov na stran

2.2 Tezave pri dostopu

- posodobitve spletne strani ter podatkovnega vmesnika, 404 večina strani.

- Pomanjkljiva dokumentacija - polje za datum je opisano vendar ni dokumentirano kaksne so vse možne vrednosti.

- manjkajoci identifikatoriji za polja na naključnih indikatorjih. primer (država ima le ime in prazno id polje) - datum vsebuje naključne vrednosti (last known value 2001 - 2015, 2040) - nedokumentira meja števila izbranih lokacij 250 - slabo definirano delovanje filtrov podatkov (mrv in date in fill) - Nemogoče ugotoviti ali ima indikator letne, cetrtlente ali mesecne vrednosti.

CCCCC

Poglavje 3

Knjižnica in gradniki za Orange

V okviru diplomske naloge smo razvili tri ločene komponente za programerje in končne uporabnike. Prva komponenta je programska knjižnica `simple_wbd`, ki omogoča enostaven dostop do programskega vmesnika indikatorjev in klimatskih podatkov Svetovne banke. Ta knjižnica je narejena s čim manj odvisnosti (dependencies, requirements?) in je namenjena splošni uporabi v python programih. Poudarek na knjižnici `simple_wbd` je predvsem enostavnost razširitve in zanesljivost. Drugi sestavni del, je razširitev knjižnice `simple_wbd` s funkcionalnostmi potrebnimi za lažje delo v programu Orange. To predvsem zavzema pretvorbo pridobljenih podatkov v Orange in numpy tabele. Ta sklop je namenjen skriptnemu delu s programom Orange in je dostopen kot `api_wrapper` python modul. Tretji sestavni del pa je grafični vmesnik za uporabo `api_wrapper` modula. Namen grafičnega vmesnika je omogočiti ne programerjem dostop do podatkov programskega vmesnika Svetovne banke znotraj programa za analizo in

3.1 Knjiznica `simple_wbd`

`Simple_wbd` je knjiznica, ki programerjem olajša dostop do podatkov programskega vmesnika svetovne banke. Glavna lastnost te knjiznice je združevanje večjega števila zahtev po podatkih, in enostavna predstavitev do-

bljenih podatkov. Druga lastnost je pretvorba podatkov iz več dimenzij, v dvo-dimenzionalno polje, ki je primerno za uporabo v programu Orange. Dva glavna vmesnika te knjižnice sta IndicatorAPI in ClimateAPI. Prvi omogoča pridobivanje podatkov z indikator programskega vmesnika Svetovne banke, drugi pa s climate programskega vmesnika.

- omogoča pridobivanje vrednosti za filtre:
 - indicator api: države in agregati, indikatorji
 - climate api: države, tipi podatkov, meritveno obdobje
- Zahteva za podatke vraca dataset objekt ki ponuja surove podatke, ali pa eno drugo obliko. 2D array ali pa dict.
- Dataset razred lahko tudi poljubno razsirimo.

3.1.1 Pomčnik IndicatorAPI

IndicatorAPI je razred namenjen pridobivanju podatkov indikatorjev razvoja. Vrednosti teh podatkov so (dolocene/ oznake s katerimi lahko do vrednosti dostopamo) z državo, casovno komponento, in imenom indikatorja. Te podatke lahko predstavimo na dva glavna načina:

- kot gnezdeni slovar, kjer je na prvem nivoju ime indikatorja, na drugem država, in na tretjem nivoju casovna komponenta.
 - Kot dvo-dimenzionalno polje, kjer imamo v vrsticah eno oznako, v stolpcih pa kartezični produkt ostalih dveh. ponujene možnosti so:
 - vrstice = država, stolpci = cas x indikator
 - vrstice = cas, stolpci = država x indikator
- indicator response:

```
[
  —
    "page": 1,
    "pages": 1,
    "per`page": "50",
    "total": 1
  ],
```

```
[
    -
    "indicator": -
        "id": "PA.NUS.PPP.05",
    "value": "2005 PPP conversion factor, GDP (LCU per international "$)"
    ,
    "country": -
        "id": "SI",
        "value": "Slovenia"
    ,
    "value": "0.611602841",
    "decimal": "2",
    "date": "2005"
    "
]
]
```

3.1.2 Pomčnik ClimateAPI

IndicatorAPI je

3.2 api_wrapper

razsiritev simple wbd vmesnikov z dedovanjem pravega dataset razreda.

```
class ClimateDataset(simple`wbd.ClimateDataset):
```

```
    def as`numpy(self):
        raise NotImplemented()
```

```
    def as`orange`table(self):
```

```
raise NotImplemented()
```

```
class ClimateAPI(simplewbd.ClimateAPI):
```

```
    def ``init``(self):  
        super().``init``(ClimateDataset)
```

3.3 Grafični vmesnik

- Lazja uporaba. - Vecja preglednost, - lazje iskanje (text search filter)

Poglavje 4

Uvod

uporaba: slike orange uporabe

4.1 Spletni viri

CCCCC

Poglavje 5

Sklepne ugotovitve

prikazuje slika 5.1.



Slika 5.1: Rainbow dash

Literatura

- [1] Jure Dimec (2002), Medjezično iskanje dokumentov
<http://clir.craynaud.com/clir/MEDJEZICNOISKANJEDOKUMENTOV.pdf>
- [2] (Avgust, 2013) OpenGL Overview
<http://www.opengl.org/about/>
- [3] (Junij, 2013) Direct3D
[http://msdn.microsoft.com/en-us/library/windows/desktop/hh309466\(v=vs.85\).aspx](http://msdn.microsoft.com/en-us/library/windows/desktop/hh309466(v=vs.85).aspx)
- [4] (Avgust, 2013) Android
<http://www.android.com/>
- [5] (Avgust, 2013) iOS
<http://www.apple.com/ios/>
- [6] (Avgust, 2013) Windows Phone
<http://developer.windowsphone.com/en-us>
- [7] (Junij, 2013) ANGLE
<https://code.google.com/p/angleproject/>
- [8] (Avgust, 2013) Firefox OS
<http://www.mozilla.org/en-US/firefox/os/>
- [9] (Avgust, 2013) Ubuntu for phones
<http://www.ubuntu.com/phone>

- [10] (Avgust, 2013) Sailfish OS
<https://sailfishos.org/>
- [11] (Avgust, 2013) Raspberry Pi FAQ
<http://www.raspberrypi.org/faqs>
- [12] (Avgust, 2013) Beagle Bone
<http://beagleboard.org/bone>
- [13] (Avgust, 2013) HTML Canvas 2D Context
<http://www.w3.org/wiki/HTML/Elements/canvas>
- [14] (Avgust, 2013) OpenGL ES 2.0 for the Web
<http://www.khronos.org/webgl/>
- [15] Vladimir Vukićević (November, 2007) Canvas 3D: GL power, web-style
<http://blog.vlad1.com/2007/11/26/canvas-3d-gl-power-web-style/>
- [16] (Avgust, 2013) SoundJS
<http://www.createjs.com/#!/SoundJS>
- [17] Michael Mahemoff (Junij, 2011), HTML5 vs Native: The Mobile App Debate
<http://www.html5rocks.com/en/mobile/nativedebate/>
- [18] (Avgust, 2013) V8-GL
<https://github.com/philogb/v8-gl>
- [19] (Avgust, 2013) LeechyJS
<http://martens.ms/lycheeJS/>
- [20] (Avgust, 2013) Xamarin
<http://xamarin.com/>
- [21] (Avgust, 2013) LibGDX
<http://libgdx.badlogicgames.com/>

-
- [22] (Februar, 2013) IKVM.NET
<http://www.ikvm.net/>
 - [23] (Avgust, 2013) PlayN
<http://code.google.com/p/playn/>
 - [24] (Avgust, 2013) Unity - Game engine, tools and mulitplatform
<http://unity3d.com/unity/>
 - [25] (Avgust, 2013) Haxe Documentation
<http://haxe.org/>
 - [26] (Avgust, 2013) Gameplay Overview
<http://www.gameplay3d.org/>
 - [27] (Avgust, 2013) OGRE - Open Source 3D Graphics Engine
<http://www.ogre3d.org>
 - [28] (Avgust, 2013) Marmalade
<http://www.madewithmarmalade.com/>
 - [29] (Avgust, 2013) Adobe Flash
<http://www.adobe.com/products/flash.html>
 - [30] (Avgust, 2013) asm.js
<http://asmjs.org/>
 - [31] (Avgust, 2013) What is CUDA
<https://developer.nvidia.com/what-cuda>
 - [32] Jonah Alben (Julij, 2013), CUDA na mobilnih napravah
<http://blogs.nvidia.com/blog/2013/07/24/kepler-to-mobile/>
 - [33] (Avgust, 2013) The open standard for parallel programming of heterogeneous systems
<http://www.khronos.org/opencv/>

[34] (Avgust, 2013) AccelerEyes

<http://www.accelereyes.com/products/mobile>

[35] (Julij, 2013) Compute Shader Overview

<http://msdn.microsoft.com/en-us/library/ff476331%28v=VS.85%29.aspx>

[36] (Avgust, 2013) LWJGL Lightweight Java Game Library

<http://www.lwjgl.org/>