SOIT

Spoločnosť pre Otvorené Informačné Technológie

# Open Scraper Challenge 2011

# Wat?

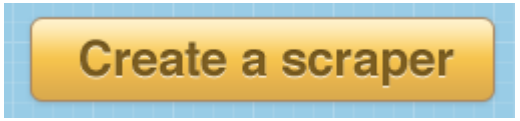
Cieľom je získať čo najviac datasetov zo Slovenska a Česka pre neskoršie použitie.

Hackovat sa bude v PHP, Python a Ruby na scraperwiki.com.

# Ideme nato: formalitky

- Join #**soit** @ irc.freenode.org
- Vytvor ucet na scraperwiki.com

# Ideme nato: process

- Vybrat dataset z Scraper Listu
- Pridat svoje meno
- 

- A scriptujeme!
- Potom otagujeme „**soit**“

# Ideme nato: priklad „Zoznam ISPs“

Limit 20

| | Poskytovateľ | F | R | M | S | MMDS | RTV | KDS | RozhlTV | VTS | PO | PD | I | VoIP | Ret | AT | Iné | Dát zač |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 111, s.r.o., Dolné Rudiny 3, 01001 Žilina | * | * | | | | | | | | | * | * | | | | | 2006 Ja |
| 2 | 3C s.r.o., Mlynská 797, 95115 Mojmírovce | | | | | | | * | | | * | | | | * | | | 1994 Jú |
| 3 | 3LOG s.r.o., Tolstého 84/8, 07901 Veľké Kapušany | * | * | | | | | | | | | | * | | | | | 2010 Fe |
| 4 | 3o media,s.r.o., , 02732 Habovka 266 | | * | | | | | * | | | | | * | * | * | | | 2007 M |
| 5 | 4CEO s.r.o., Gorlická 5, 08501 Bardejov | | * | | | | | | | | | | * | | | | | 2006 Jú |
| 6 | A&J, s.r.o., Šafárikova 2735/1, 91108 Trenčín | | * | | | | | | | | | * | * | | | | | 2010 Ja |
| 7 | A-Network s.r.o., Brezová 75, 90023 Viničné pri Pezinku | | | | | | | | | | | | * | | | | | 2004 Au |
| 8 | A.I.S., s.r.o., Staré Grunty 53, 84104 Bratislava | | * | | | | | | | | | | * | | | | | 2002 Ja |
| 9 | Abiks s.r.o., Kollárova 1304/14, 01841 Dubnica n. Váhom | | * | | | | | | | | | | * | | | | | 2004 M |
| 10 | ACS, s.r.o., Ružová dolina 10, 82109 Bratislava | * | * | | | | | | | * | | | * | | | | | 2005 Se |
| 11 | Active World Business s.r.o., Tajovského 1, 96301 Krupina | | * | | | | | | | | | * | * | | | | | 2008 Ja |
| 12 | AD service, spol. s r.o., Strojárska 362, 96601 Hliník nad Hronom | | | | | | | | | | * | * | * | * | | | | 2009 Jú |

# Ideme nato: priklad Script

**Pavol Rusnak / Slovak Providers of Networks and Services**

```
1  import scraperwiki
2  import lxml.html
3  import time
4
5  # workaround of non-working sk SK locale
6  # see https://bitbucket.org/ScraperWiki/scraperwiki/issue/526/python-mis
7  def fix_month(text):
8      if isinstance(text, str):
9          return text.replace('Máj', 'May').replace('Jún','Jun').replace('
10     else:
11         return text
12
13
14 html = scraperwiki.scrape('http://www.teleoff.gov.sk/sk/OTR/viewpublic.p
15 html = html.replace('\x00', ' ') # fix broken html :-)
16 root = lxml.html.fromstring(html)
17
18 for tr in root.cssselect("table[class='fotky'] tr"):
19     tds = tr.cssselect("td")
20     if len(tds) < 23:
21         continue
22     try:
23         start_date = int(time.mktime(time.strptime(fix_month(tds[18].tex
24     except:
25         start_date = None
```

# Ideme nato: priklad dataset

swdata (1172 rows)

Download

| comment | address | rtv | radio | pts | id | end | retransmission | start | catv | int |
|---|---|---|---|---|---|---|---|---|---|---|
| internet od 1.4.2005, VoIP od 1. 5. 2008 | SR | 0 | 0 | 1 | 1170 | | 0 | 1038700800 | 0 | 1 |
| WiFI sieť; Prešovský kraj - Kežmarok, Spišská Belá | SR | 0 | 1 | 0 | 1169 | | 0 | 1298937600 | 0 | 0 |
| rádiová sieť v pásme 2,4 a 5 GHz, | SR | 0 | 1 | 0 | 1168 | | 0 | 1283299200 | 0 | 1 |
| WiFi sieť v pásme 2,4 GHz a 5470 - 5725 MHz | západné Slovensko | 0 | 1 | 0 | 1167 | | 0 | 1220227200 | 0 | 1 |
| Poskytovanie elektronických služieb v Ciernom Balogu od 1.1.2006.,WIFI sieť v pásme 2,4 GHz | Podbrezová Štiavnička, Valaska,Cierny Balog | 0 | 1 | 0 | 1166 | | 0 | 1101859200 | 0 | 1 |
| Frelvenčné pásmo 2,4 a 5 | okres Rimavská | 0 | 1 | 0 | 1165 | | 0 | 1312156800 | 0 | 1 |

# Otazky?

# Credits

- http://www.flickr.com/photos/tambako/4216369204/
- http://www.flickr.com/photos/zeevveez/4668099379/