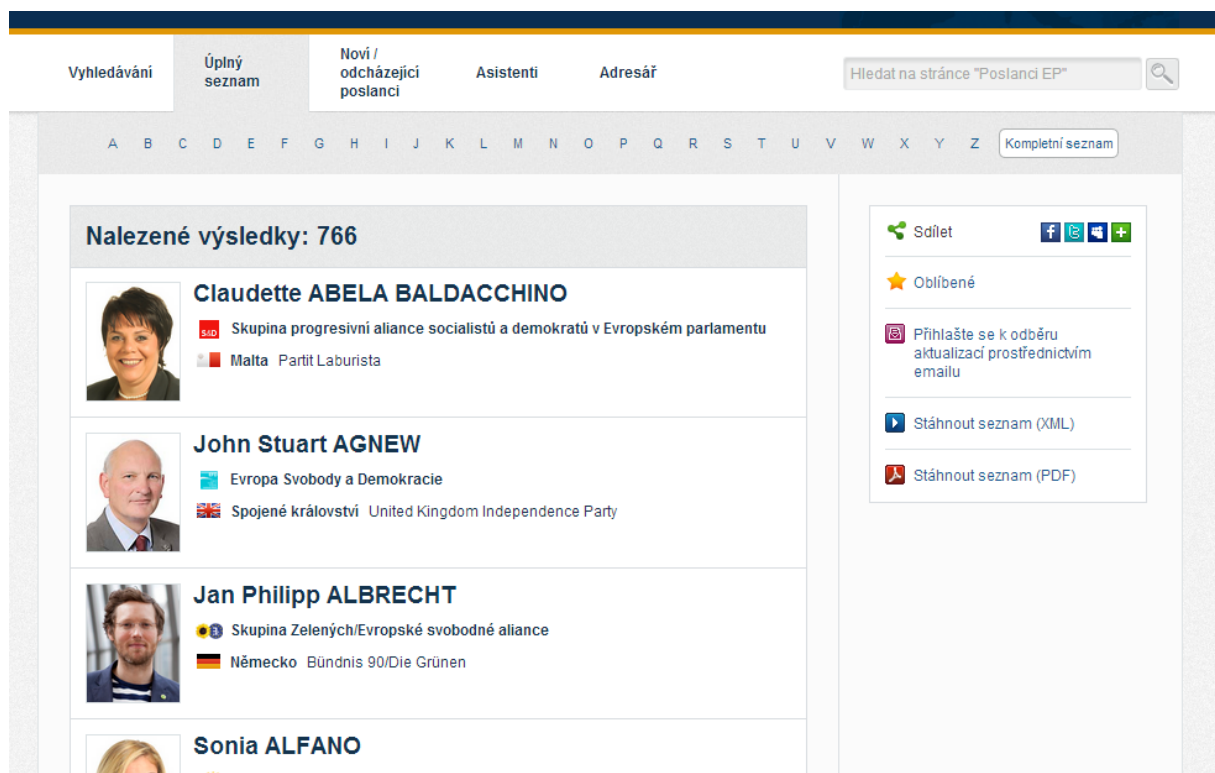


Ve zcela nepovinném pseudoprogramovacím domácím cvičení si vyzkoušíme velmi jednoduché scrapování jednoduchého webu. Jak ukazoval na čtvrtěční přednášce Honza Boček, scrapování pomáhá tam, kde data existují, jsou přístupná, ale nejsou ve vhodných podobách. Scrapování není tak jednoduché, weby jsou někdy složité, musíte pracovat s různými strukturami URL (vzpomeňte atlety v Sochi, kde výška a váha byla na samostatné stránce každého atleta, atp.) My si zkusíme jednodušší web, kde jsou všechny informace, které potřebujeme, na jedné stránce.

Modelová situace: plánujete vizualizaci o Evropském parlamentu a jako výchozí krok potřebujete tabulku, která bude obsahovat jména všech poslanců, jejich národnost a stranickou příslušnost. Předpokládejme nyní, že taková tabulka nikde předpřipravená není.

Na stránce Evropského parlamentu jste našli seznam všech poslanců, je na jedné stránce, včetně požadovaných informací. Vypadá přesně takhle:



<http://www.europarl.europa.eu/meps/cs/full-list.html?filter=all&leg=>

766 poslanců tu visí pod sebou a tabulku nevidno. Prostě Ctrl+C a Ctrl+V nefunguje a ruční překopírovávání do sloupečků by trvalo dlouhé dny. Přesně tady přichází na řadu scrapování.

Jelikož v oblasti nástrojů pro práci s daty a pro vizualizaci došlo k obrovské laicizaci (*i o tom budeme mluvit příští čtvrtek*), nemusíte dnes umět nijak výrazně programovat, abyste byli schopni scrapovat takto jednoduchý web. Jeden z přístupných nástrojů pro scrapování si dnes ukážeme.

Nainstalujte si do Chrome doplněk Scraper.

Zamířte pro něj na adresu:

<https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecacpepngjd>

Scraper je jednoduchý prográmeček, který používá jazyk XPath (<http://cs.wikipedia.org/wiki/XPath>), sloužící, zjednodušeně řečeno, k odkazování na strukturu dokumentu. Na bakaláři jsme měli HTML i XML, takže si pamatujete, že uvnitř to všechno vypadá nějak takhle:

```
<produkt>
  <polozka typ="technika">auto</polozka>
  <polozka typ="technika">kolo</polozka>
  <polozka typ="obleceni">boty</polozka>
</produkt>

<div class="odkaz">
  <a href="http://www.google.com">Google</a>
  <span class="typ_zdroje">Vyhledávač</span>
</div>
```

Prostě to má hierarchickou strukturu. Pomocí XPath můžete Scraperu např. říct, ať vám z XML dokumentu obsahujícího seznam produktů vytáhne jen a pouze ty produkty, které jsou typu „elektronika“ a hodí je pod sebe do tabulky.

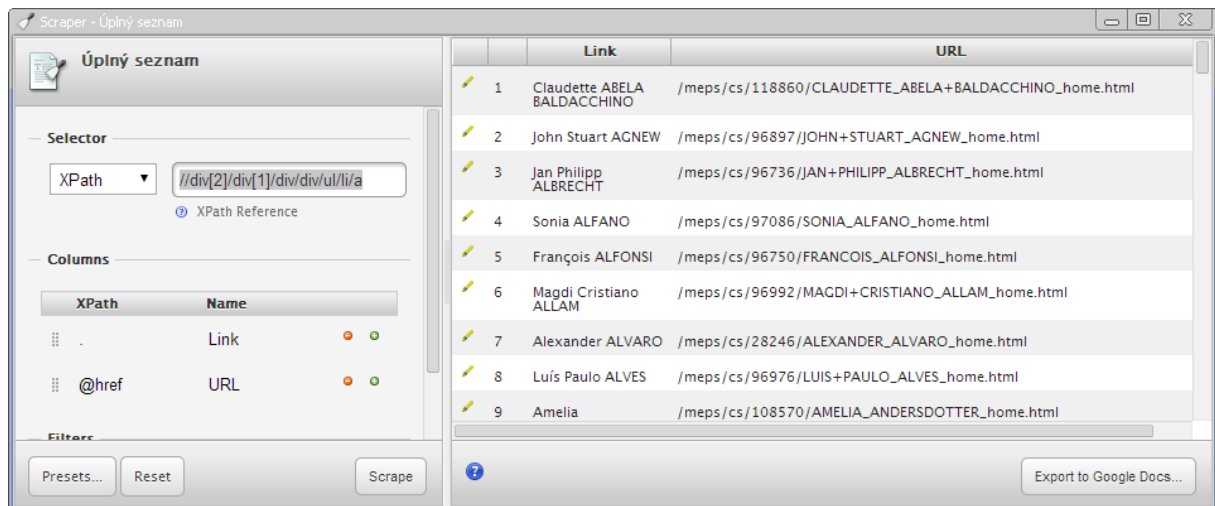
Přesně to teď uděláme na stránce s kompletním seznamem Evropských poslanců. Řekneme skrze základní příkazy XPath Scraperu, aby nám z webu vytáhl jméno poslance, jeho národnost a stránickou příslušnost a aby to takhle udělal u všech 766 poslanců a aby nám z toho nakonec udělal i hezkou tabulku.

Klikněte pravým tlačítkem na jméno první poslankyně a zvolte *Zkontrolovat prvek*. Rozbalí se vám náhled HTML naší stránky.

The screenshot shows a search result for 'Claudette ABELA BALDACCHINO' from Malta, a member of the European Parliament. The browser's developer tools are open, displaying the HTML structure. The selected element is an `li` with class `mep_name`, containing a link to the member's profile and their political affiliation: 'Skupina progresivní aliance socialistů a demokratů v Evropském parlamentu'.

Prozkoumejte si pečlivě kód. Jméno poslankyně je v HTML uloženo v tagu `<a>` (odkaz na podstránku) a především pak v tagu ``, což je položka seznamu. Tento tag `` je dále uložen spolu s dalšími `` v tagu ``, který v HTML uvozuje odrážkový seznam. Celý odrážkový seznam je pak uložen uvnitř tagu `<div>`, který je uložen uvnitř dalšího `<div>` a ten, pro změnu, uvnitř dalšího `<div>`. Struktura dále pokračuje, ale my už ji zkoumat nebudeme. Důležité je pochopit, že každý prvek – jméno, název stranz, národnost – má v rámci dokumentu definovatelnou cestu.

Klikněte na jméno poslankyně Claudette pravým tlačítkem znovu a zvolte *Scrape similar...*



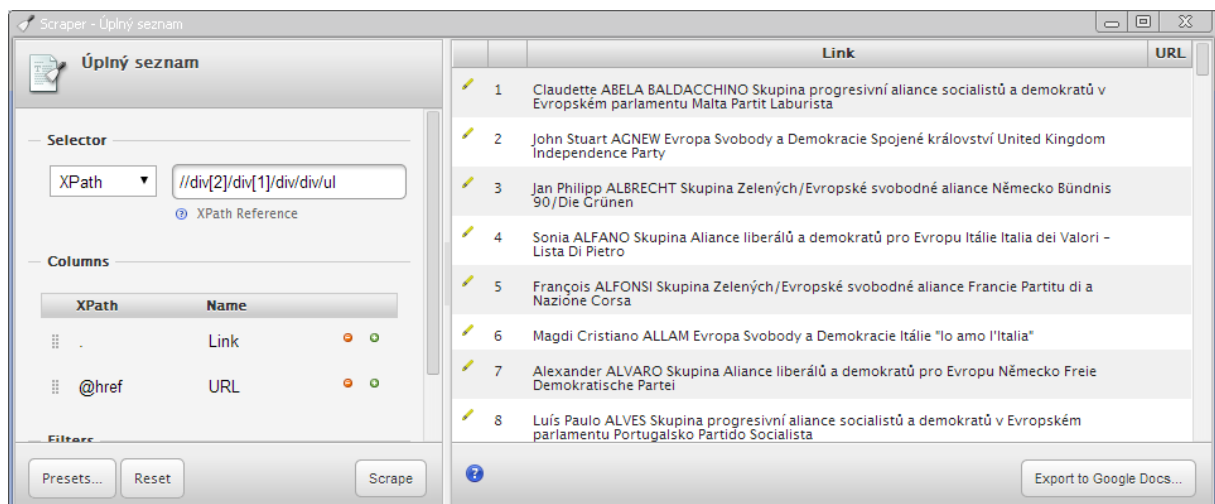
Náš **Scraper** právě identifikoval strukturu dokumentu (i bez našeho výrazného snažení), našel si v HTML tazích cestu ke jménu poslankyně a také podobné cesty ke jménům dalších poslanců. V okně vlevo vidíte příkaz XPath, pod ním pak „podpříkazy“, které určují jaké další informace se z daného umístění do sloupců tabulky načítají – k tomu se ještě dostaneme.

V okně napravo pak vidíte výsledek scrapování – tabulka se 766 jmény a URL odkazy na podstránky pro jednotlivé poslance. Z údajů, které reálně potřebujeme, ale nyní máme jen jméno.

Scraper identifikoval jako XPath selector „adresu“ `//div[2]/div[1]/div/div/ul/li/a`

My ale z našeho zkoumání struktury víme, že tady se nachází pouze jméno a odkaz na podstránku – další potřebné údaje se nacházejí jinde. Osekáme tedy adresu až na ten tag, který je společným nadřazeným všem tagům, jež obsahují námi chtěné informace. To zní trochu chaoticky, takže raději hned prakticky.

Zkuste z adresy vymazat poslední `/a` – zůstane tedy `//div[2]/div[1]/div/div/ul/li` a klikněte na tlačítko Scrape. To není ono, stále nejsme dost vysoko. Umažte tedy ještě `/li` – zůstane `//div[2]/div[1]/div/div/ul` a klikněte na tlačítko Scrape. To vypadá lépe. Scraper „projel“ všechny `//div[2]/div[1]/div/div/ul` v dokumentu a vytáhl z nich informace – v tabulce nyní vidíme všechny údaje, které potřebujeme: je tam jméno, je tam strana, spolek i národnost. Má to jednu chybu: všechno je to v jednom jediném sloupci.



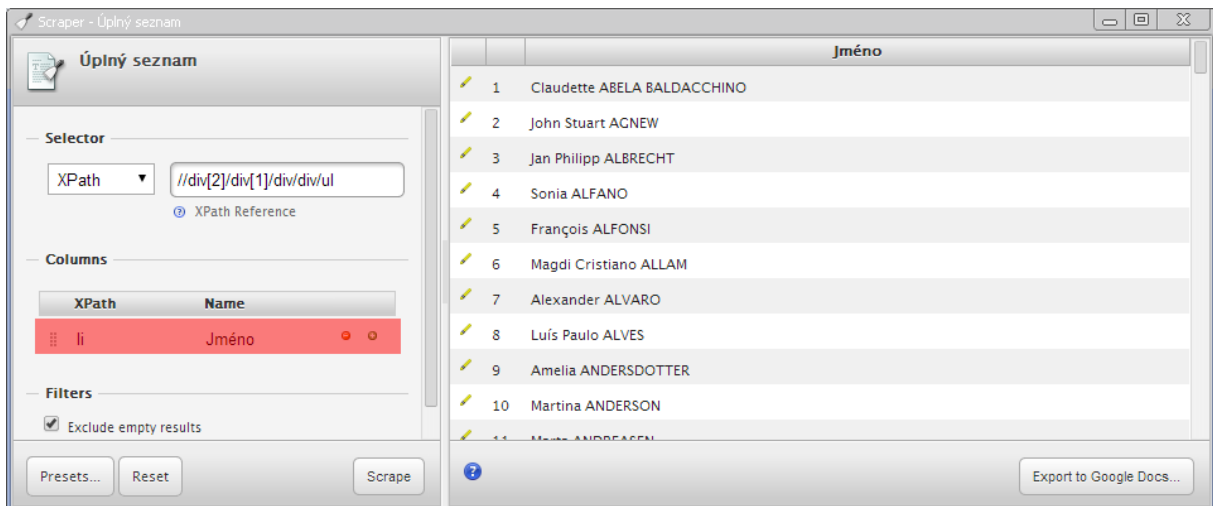
```

<div class="zone_info_mep">
  <a href="/meps/cs/118860/CLAUDETTE_ABELA+BALDACCHINO_home.html">...</a>
  <div class="mep_details">
    <ul>
      <li class="mep_name">
        <a href="/meps/cs/118860/CLAUDETTE_ABELA+BALDACCHINO_home.html" class="Claudette ABELA BALDACCHINO">
          </li>
        <li class="group sd">Skupina progresivní aliance socialistů a demokratů v Evropském parlamentu</li>
      </ul>
    </div>
  <div class="clear">&nbsp;</div>
</div>
<div class="zone_info_mep">...</div>

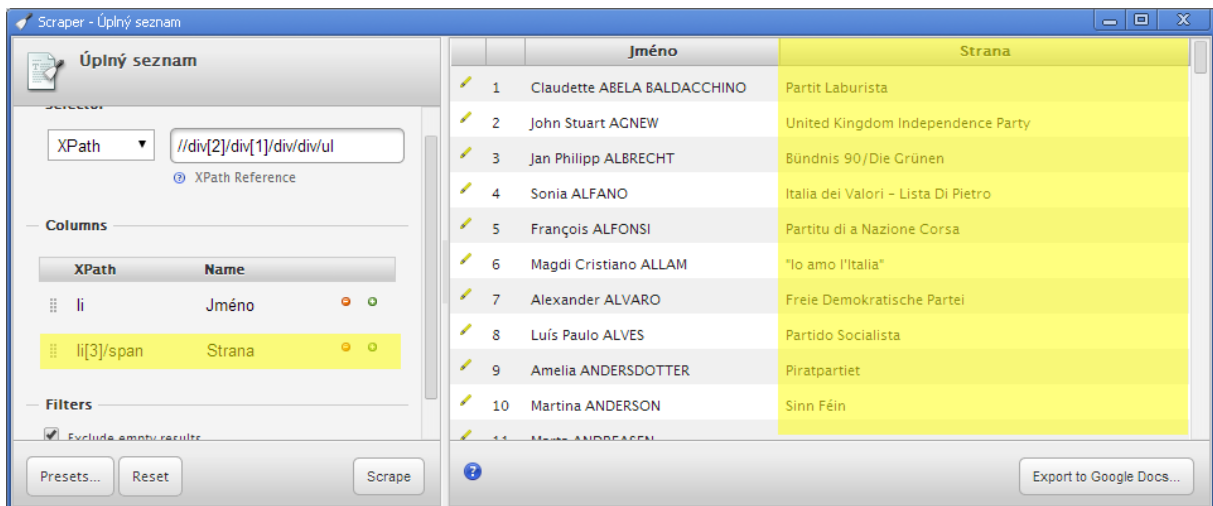
```

To samé si ukážeme ještě v HTML. Scaper nyní tedy u každého poslance dojde až k tagu ul (červeně) a vytáhne z něj všechny informace (tedy vše, co je uvnitř žlutého rámečku). Pomocí nabídky *Columns* v pravém sloupci nyní Scaperu upřesníme, jaké má z `//div[2]/div[1]/div/div/ul` tahat další „podtagy“ a do jakých sloupců je má zařadit.

Smažme tedy ty, co tam nyní jsou, a to pomocí malého červeného mínuska. V kódu vidíme, že v rámci tagu `` je jméno poslance umístěno v prvním podtagu ``. Vše co teď uděláme je, že do pole XPath napíšeme `li` a do pole *Name* sloupec pojmenujeme jako *Jméno*. Klikneme na tlačítko *Scrape*.



Máme v tabulce samostatný sloupeček pro jméno. Zpátky ke kódu. V něm vidíme, že poslancova domovská strana se nachází trochu hlouběji ve struktuře, v třetím tagu `` a dále v podtagu ``. Pomocí zeleného pluska tedy přidáme další sloupec. Nazveme ho *Strana* a jako XPath cestu zadáme `li[3]/span` a klikneme na tlačítko *Scrape*. Scaper nyní projede všechny poslance a u každého vytáhá informace z cesty, kterou jsme mu právě zadali – tedy ze třetího `[3]` tagu `` a podtagu ``.



Pokrok! Máme tabulku se jmény všech poslanců EP a jejich stranickou příslušností. Co nám zbývá? Národnost. Zpátky ke kódu.

```

<div class="mep_details">
  <ul>
    <li class="mep_name">
      <a href="/meps/cs/118860/CLAUDETTE_ABELA+BALDACCHINO_home.html" class">Claudette ABELA BALDACCHINO</a>
    </li>
    <li class="group_sd">Skupina progresivní aliance socialistů a demokratů v Evropském parlamentu</li>
    <li class="nationality_mt">
      "Malta"
      <span class="name_pol_group">Partit Laborista</span>
    </li>
  </ul>
</div>
<div class="clear">&nbsp;</div>
</div>

```

Údaj o národnosti se nachází ve třetím tagu pod tagem . Logicky tedy vyplníme **li[3]** a sloupec nazveme Národnost. Klikneme na tlačítko *Scrape* a...



...a něco není v pořádku. XPath poslušně vytáhl všechny informace z druhého tagu pod tagem , ale protože jeho součástí je i tag s názvem strany, jsou teď oba tyto údaje v jednom sloupci. Tady přichází na řadu vyšší XPath dívčí. Musíme mu nějakým způsobem říct, aby z třetího tagu vytáhl jen a pouze údaj o národnosti a to, co je tam dále v podtagu , nechal na pokoji. Scraperu to řekneme jazykem, kterému rozumí – tedy XPath: **li[3]/node()[not(self::span)]**

Tento příkaz zajistí, že si do sloupce Národnost opravdu vytáhne jen národnosti, ale všechny další podtagy nechá ležet bez povšimnutí. Zkusme XPath příkaz ve sloupci Národnost upravit na tento pokročilejší a kliknout na tlačítko *Scrape*.

The screenshot shows the Scraper tool interface. On the left, the 'Columns' section is visible, showing the configuration for the 'Národnost' column. The XPath expression is set to `li[3]/node()[not(s Národnost)]`. The main table displays the following data:

	Jméno	Strana	Národnost
1	Claudette ABELA BALDACCHINO	Partit Laborista	Malta
2	John Stuart AGNEW	United Kingdom Independence Party	Spojené království
3	Jan Philipp ALBRECHT	Bündnis 90/Die Grünen	Německo
4	Sonia ALFANO	Italia dei Valori - Lista Di Pietro	Itálie
5	François ALFONSI	Partitu di a Nazione Corsa	Francie
6	Magdi Cristiano ALLAM	"Io amo l'Italia"	Itálie
7	Alexander ALVARO	Freie Demokratische Partei	Německo
8	Luís Paulo ALVES	Partido Socialista	Portugalsko
9	Amelia ANDERSDOTTER	Piratpartiet	Švédsko
10	Martina ANDERSON	Sinn Féin	Spojené

A jsme doma! V pravém sloupci je tabulka všech poslanců EP se jmény, národností a stranickou příslušností. Jistě jste si všimli, že na stránce EP je kromě těchto údajů uvedena ještě politická skupina uvnitř parlamentu, kterou je daný poslanec či daná poslankyně členem či členkou. Zkuste si sami ve Scraperu přidat další sloupec a do něj si nechat vytáhnout název spolku. Povedlo se?

Pak už stačí jen kliknout na *Export to Google Docs* a tabulka je na světě. Gratuluji, právě jste *poscrapovali* (nebo podle Honzy „olízali“) web Evropského parlamentu!

The screenshot shows a Google Docs spreadsheet with the following data:

	Jméno	Strana	Národnost
1	Jméno	Strana	Národnost
2	Claudette ABELA BALDACCHINO	Partit Laborista	Malta
3	John Stuart AGNEW	United Kingdom Independence Party	Spojené království
4	Jan Philipp ALBRECHT	Bündnis 90/Die Grünen	Německo
5	Sonia ALFANO	Italia dei Valori - Lista Di Pietro	Itálie
6	François ALFONSI	Partitu di a Nazione Corsa	Francie
7	Magdi Cristiano ALLAM	"Io amo l'Italia"	Itálie
8	Alexander ALVARO	Freie Demokratische Partei	Německo
9	Luís Paulo ALVES	Partido Socialista	Portugalsko
10	Amelia ANDERSDOTTER	Piratpartiet	Švédsko
11	Martina ANDERSON	Sinn Féin	Spojené království
12	Marta ANDREASEN	-	Spojené království
13	Josefa ANDRÉS BAREA	Partido Socialista Obrero Español	Španělsko
14	Eric ANDRIEU	Parti socialiste	Francie
15	Laima Liucija ANDRIKIENĖ	Tėvynės sąjunga - Lietuvos krikščionys demokratai	Litva
16	Roberta ANGELILLI	Nuovo Centrodestra	Itálie
17	Charalampos ANGOURAKIS	Communist Party of Greece	Řecko
18	Antonello ANTINORO	Unione dei Democratici cristiani e dei Democratici di Centro	Itálie
19	Elena Oana ANTONESCU	Partidul Democrat-Liberal	Rumunsko
20	Alfredo ANTONIOZZI	Nuovo Centrodestra	Itálie
21	Pablo ARIAS ECHEVERRÍA	Partido Popular	Španělsko
22	Pino ARLACCHI	Partito Democratico	Itálie