Session 15
Gathering Electronic Data 1

R for Social Data Science

Jeffrey Ziegler, PhD

Assistant Professor in Political Science & Data Science
Trinity College Dublin

Fall 2022

Road map for today

Last week:

- Data visualization

This time:

- Online data sources
- Data collection
- Web technologies
- HTML fundamentals

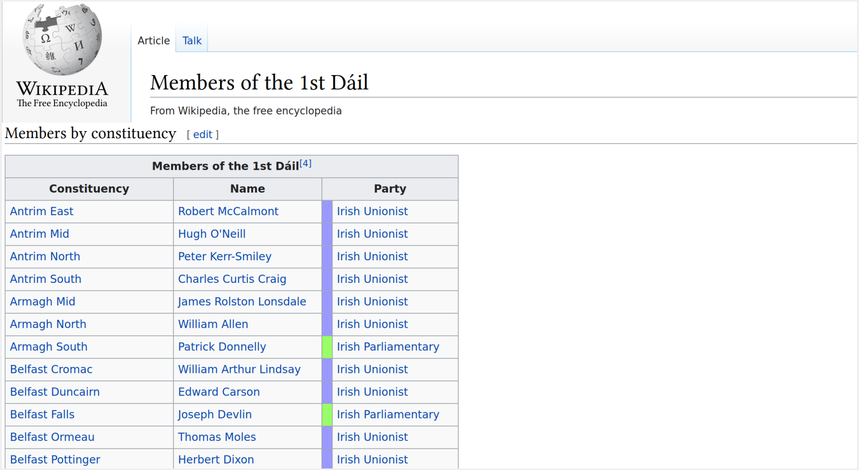
Online data sources

- Data downloadable in tabular format (E.g. CSV/TSV, XLS, DTA, etc.)
- Data available online as a table (E.g. webpages with rendered tables)
- Unstructured data available online (E.g. simple webpages)
- Interactive webpages with user-input (E.g. webpages with logins, dropdown menus)
- Web APIs (special interfaces for querying, e.g. Twitter, Google)

Online data collection

- Tabular format: download single or multiple files (automate with ‘download.file()‘ in R, ‘wget‘ in Python/Terminal)
- Online tables and unstructured data: simple web scraping (HTML with XPath, ‘rvest‘ in R, ‘beautifulsoup‘ in Python)
- Interactive webpages: web scraping with headless browser (Selenium, ‘RSelenium‘ in R, ‘selenium‘ in Python)
- Web API: sending requests and processing responses (HTTP queries, ‘httr‘ in R, ‘requests‘ in Python)

Web tables



Members of the 1st Dáil

From Wikipedia, the free encyclopedia

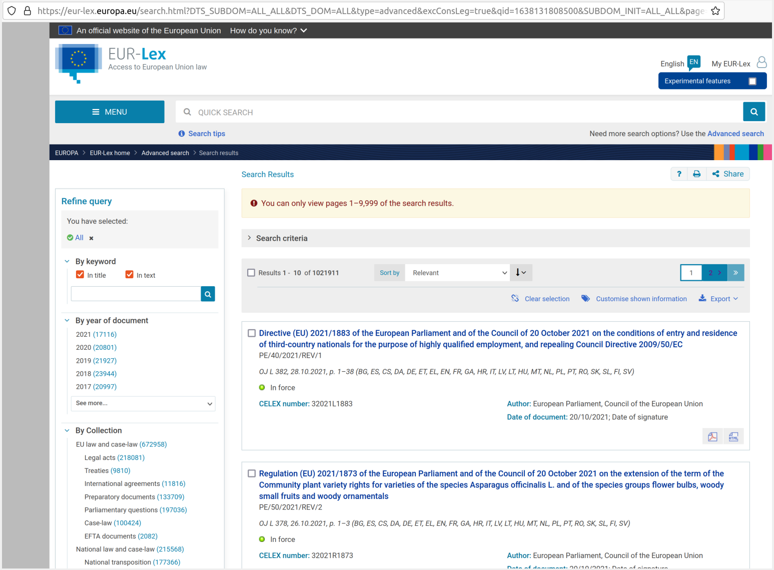
Members by constituency [edit]

Members of the 1st Dáil[4]

Constituency	Name	Party	
Antrim East	Robert McCalmont	🟦	Irish Unionist
Antrim Mid	Hugh O'Neill	🟦	Irish Unionist
Antrim North	Peter Kerr-Smiley	🟦	Irish Unionist
Antrim South	Charles Curtis Craig	🟦	Irish Unionist
Armagh Mid	James Rolston Lonsdale	🟦	Irish Unionist
Armagh North	William Allen	🟦	Irish Unionist
Armagh South	Patrick Donnelly	🟩	Irish Parliamentary
Belfast Cromac	William Arthur Lindsay	🟦	Irish Unionist
Belfast Duncairn	Edward Carson	🟦	Irish Unionist
Belfast Falls	Joseph Devlin	🟩	Irish Parliamentary
Belfast Ormeau	Thomas Moles	🟦	Irish Unionist
Belfast Pottinger	Herbert Dixon	🟦	Irish Unionist

Source: Wikipedia

Unstructured data



https://eur-lex.europa.eu/search.html?DTS_SUBDOM=ALL_ALL&DTS_DOM=ALL&type=advanced&excConsLeg=true&qid=1638131808500&SUBDOM_INIT=ALL_ALL&page

An official website of the European Union How do you know?

EUR-Lex
Access to European Union law

English EN My EUR-Lex

Experimental features □

MENU

QUICK SEARCH

Search tips

Need more search options? Use the Advanced search

EUROPA > EUR-Lex home > Advanced search > Search results

Search Results

? Share

Refine query

You have selected:

All ×

By keyword

☑ In title ☑ In text

By year of document

2021 (17116)
2020 (20801)
2019 (21927)
2018 (23944)
2017 (20997)

See more...

By Collection

EU law and case-law (672958)
Legal acts (218081)
Treaties (9810)
International agreements (11816)
Preparatory documents (133709)
Parliamentary questions (197036)
Case-law (100424)
EFTA documents (2082)
National law and case-law (215568)
National transposition (177366)

You can only view pages 1–9,999 of the search results.

Search criteria

□ Results 1 - 10 of 1021911

Sort by Relevant

1 2 >

Clear selection Customise shown information Export

□ **Directive (EU) 2021/1883 of the European Parliament and of the Council of 20 October 2021 on the conditions of entry and residence of third-country nationals for the purpose of highly qualified employment, and repealing Council Directive 2009/50/EC**
PE/40/2021/REV/1

OJ L 382, 28.10.2021, p. 1–38 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

● In force

CELEX number: 32021L1883

Author: European Parliament, Council of the European Union

Date of document: 20/10/2021; Date of signature

□ **Regulation (EU) 2021/1873 of the European Parliament and of the Council of 20 October 2021 on the extension of the term of the Community plant variety rights for varieties of the species Asparagus officinalis L. and of the species groups flower bulbs, woody small fruits and woody ornamentals**
PE/50/2021/REV/2

OJ L 378, 26.10.2021, p. 1–3 (BG, ES, CS, DA, DE, ET, EL, EN, FR, GA, HR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV)

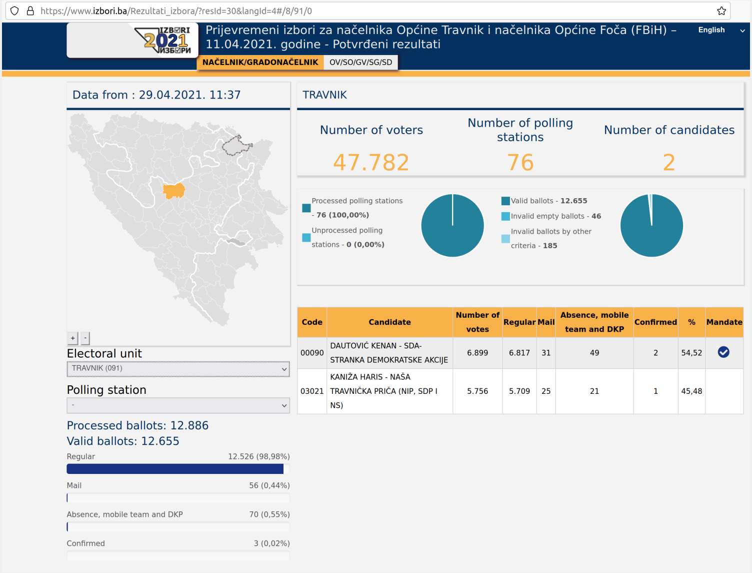
● In force

CELEX number: 32021R1873

Author: European Parliament, Council of the European Union

Source: Eur-Lex

Interactive webpages



https://www.izbori.ba/Rezultati_izbora/?resId=30&langId=4#/8/91/0

IZBORI 2021 ИЗБОРИ

Prijevremeni izbori za načelnika Općine Travnik i načelnika Općine Foča (FBiH) – 11.04.2021. godine - Potvrđeni rezultati

English

NAČELNIK/GRADONAČELNIK | OV/SO/GV/SG/SD

Data from : 29.04.2021. 11:37

TRAVNIK

Number of voters	Number of polling stations	Number of candidates
47.782	76	2

■ Processed polling stations - **76 (100,00%)**

■ Unprocessed polling stations - **0 (0,00%)**

■ Valid ballots - **12.655**

■ Invalid empty ballots - **46**

■ Invalid ballots by other criteria - **185**

Code	Candidate	Number of votes	Regular	Mail	Absence, mobile team and DKP	Confirmed	%	Mandate
00090	DAUTOVIĆ KENAN - SDA-STRANKA DEMOKRATSKE AKCIJE	6.899	6.817	31	49	2	54,52	✓
03021	KANIŽA HARIS - NAŠA TRAVNIČKA PRIČA (NIP, SDP I NS)	5.756	5.709	25	21	1	45,48	

Electoral unit: TRAVNIK (091)

Polling station: -

Processed ballots: 12.886

Valid ballots: 12.655

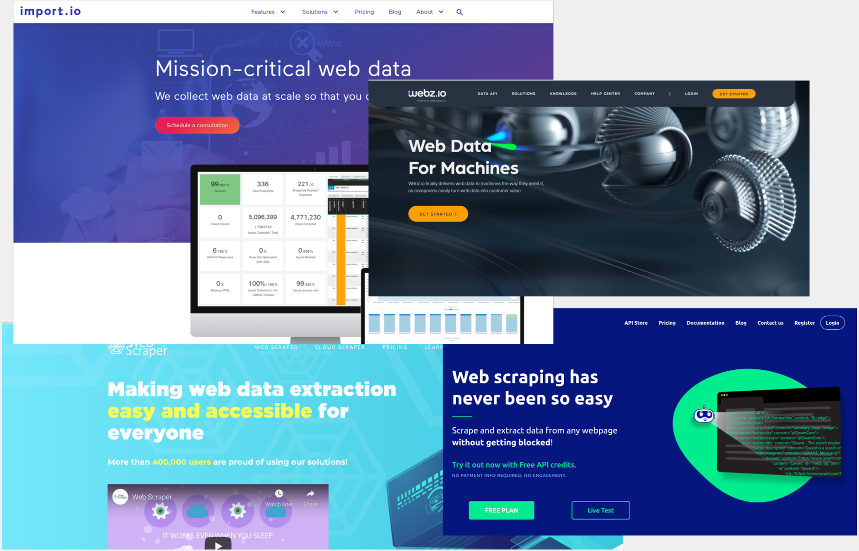
Regular	12.526 (98,98%)
Mail	56 (0,44%)
Absence, mobile team and DKP	70 (0,55%)
Confirmed	3 (0,02%)

Source: Izbori.ba

Automated data collection

- Manual scraping (copy-pasting) can be:
    - Extremely laborious and time-consuming
    - Very error-prone
    - Often impossible to reproduce exactly
    - Automated data collection
    - Easy to scale up (computer time is cheap)
    - Less error-prone
    - Usually, perfectly reproducible
    - There is a trade-off (time invested in automation vs time saved)
    - However, it is good to err on the side of automation

Commercial solutions



import.io

Features Solutions Pricing Blog About

Mission-critical web data

We collect web data at scale so that you c

Schedule a consultation

99.987% Success	336 Total Snapshots	221/0 Snapshots Pushed / Expected
0 Failed Imports	5,096,399 / 7382733 Inputs Collected / Total	4,771,230 Rows Extracted
6.192% 404/410 Responses	0% Rows Not Generated (with 200)	0.009% Inputs Blocked
0% Missing HTML	100% / 100% Inputs Collected in 1hr / Before Timeout	99.433% Inputs process rate

webz.io
Formerly Webhose.io

DATA API SOLUTIONS KNOWLEDGE HELP CENTER COMPANY | LOGIN GET STARTED

Web Data For Machines

Webz.io finally delivers web data to machines the way they need it, so companies easily turn web data into customer value

GET STARTED ›

Web Scraper

WEB SCRAPER CLOUD SCRAPER PRICING LEARN

Making web data extraction easy and accessible for everyone

More than 400,000 users are proud of using our solutions!

Web Scraper Watch later Share

IT WORKS EVEN WHEN YOU SLEEP

API Store Pricing Documentation Blog Contact us Register Login

Web scraping has never been so easy

Scrape and extract data from any webpage **without getting blocked**!

Try it out now with Free API credits.

NO PAYMENT INFO REQUIRED, NO ENGAGEMENT.

FREE PLAN Live Test

```
charset="utf-8">
http-equiv="X-UA-Compatible" content="IE=edge">
<title>Qwant</title>
name="twitter:card" content="summary_large_image">
name="twitter:site" content="@QwantCom">
name="twitter:creator" content="@QwantCom">
name="twitter:title" content="Qwant - The search engine
"twitter:description" content="Qwant is a search er
itter:image:src" content="/QWANT_Meta.png?1
domain" content="https://www.qwant.com
" content="Qwant" id="meta_og_title"/
e" content="Qwant"/>
nt="https://www.qwant.com/
```

Web technologies

- Key technologies used to disseminate content on the Web:
  - XML/HTML (E**x**tensible **M**arkup **L**anguage/**H**yper**t**ext **M**arkup **L**anguage)
  - CSS (**C**ascading **S**tyle **S**heets)
  - JavaScript
  - API (**A**pplication **P**rogramming **I**nterface)
  - JSON (**J**ava**S**cript **O**bject **N**otation)

Static vs dynamic websites

- The critical feature of a website which determines approach to scraping its content
- Static websites all have prebuild source code which is served at user's request
  - No real-time processing of user's input
  - Can contain elements that change appearance of a website
  - Example: Course website
  - Dynamic websites render websites in real-time as a response to user's input
  - They can use a range of technologies to achieve it (JavaScript, Python Django, PHP)
  - Example: Google Maps

HTML: Hypertext Markup Language

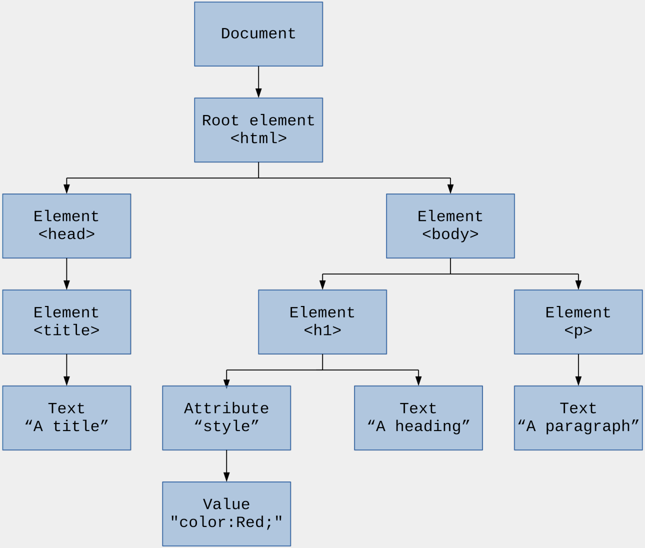
- HTML (**H**yper**t**ext **M**arkup **L**anguage) is a mark-up language for webpages
- Forms the basis of static websites
- Your browser renders (interprets) HTML for viewing
- Current version is HTML5

Extra - W3Schools: Try HTML

HTML BASICS

- Basic unit of HTML is an *element* (aka *node*)
- Elements, typically, begin with an start tag (e.g. ‘<h1>‘)
- And finish with an *end tag* (e.g. ‘</h1>‘)
- Content of element is found between start and end tags
- *Attributes* are special words used within a start tag to control element’s behaviour (e.g. ‘style="color:Red;"‘)
- Some HTML tag examples:
  - Document structure: ‘<html>‘, ‘<body>‘, ‘<header>‘
  - Document components: ‘<h1>‘, ‘<title>‘, ‘<div>‘
  - Text style: ‘<b>‘, ‘<i>‘
  - Hyperlinks: ‘<a>‘

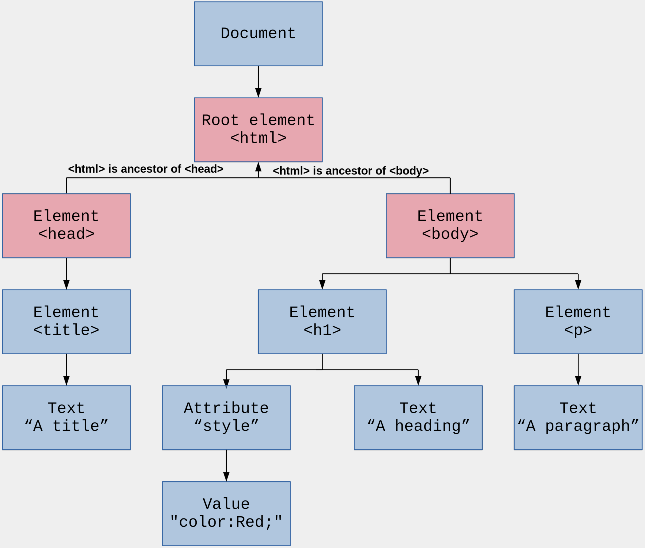
HTML TREE


Document
Root element
<html>
Element
<head>
Element
<body>
Element
<title>
Element
<h1>
Element
<p>
Text
“A title”
Attribute
“style”
Text
“A heading”
Text
“A paragraph”
Value
"color:Red;"

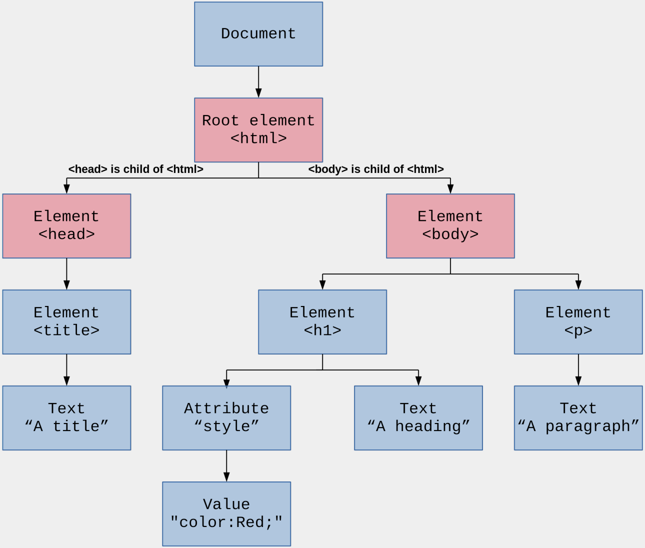
HTML tree relationships

- All elements (nodes) in HTML tree are connected by relationships
- These relationship can be of the following types:
  - Ancestors (parents)
  - Descendants (children)
  - Siblings

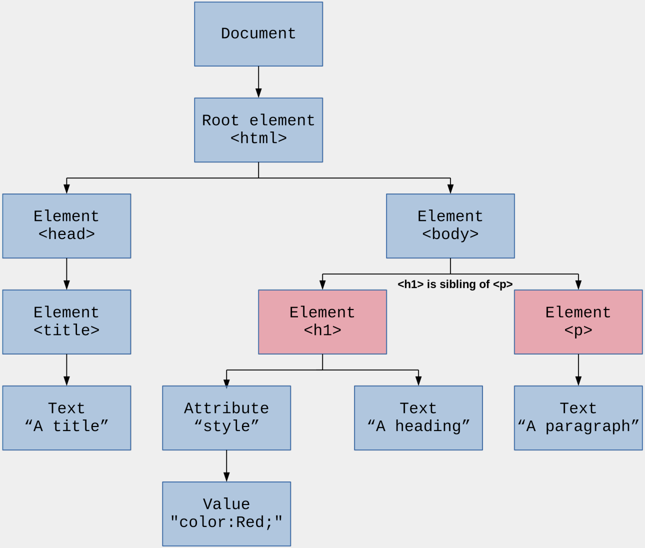
HTML PARENT/ANCESTOR



HTML CHILDREN/DESCENDANTS



HTML SIBLINGS



Ex: Parsing HTML tree

```r
html_txt <- "\n
<!DOCTYPE html> \n
<html>\n
    <head>\n
        <title>A title</title> \n
    </head>\n
    <body>\n
        <h1 style=’color:Red;’>A heading</h1> \n
        <p>A paragraph.</p> \n
    </body>\n
</html>"
html <- rvest::read_html(html_txt)
str(html)
```

```
List of 2
$ node:<externalptr>
$ doc :<externalptr>
- attr(, "class")= chr [1:2] "xml_document" "xml_node"
```

Ex: Parsing HTML tree

```r
children <- rvest::html_children(html)
children
```

```
{xml_nodeset (2)}
[1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n<title>A title
[2] <body>\n\n        <h1 style="color:Red;">A heading</h1> \n\n        <p>A paragraph.</p> \n\
```

```r
body <- children[2]
rvest::html_name(body)
```

```
[1] "body"
```

Ex: Parsing HTML tree

```r
children2 <- rvest::html_children(body)
children2
```

```
{xml_nodeset (2)}
[1] <h1 style="color:Red;">A heading</h1>
[2] <p>A paragraph.</p>
```

```r
rvest::html_attrs(children2[1])
```

```
[[1]]
style
"color:Red;"
```

```r
rvest::html_text(children2[1])
```

```
[1] "A heading"
```

Tutorial - HTML basics and scraping tables

- We will extract the table of countries with their GDP from a Wikipedia article
- Start by loading in the webpage using ‘rvest“s ‘read_html()‘ function
- Go the webpage of the article and locate the elements that would be helpful for table extraction
- Extract the ‘<table>‘ node that correponds to the main table
- Extract ‘<tbody>‘ element as a child of this element
- Extract the table of with data using ‘rvest“s ‘html_table()‘ function
- Tidy up the extracted table

Overview

This time:

- Online data sources
- Data collection
- Web technologies
- HTML fundamentals

Next time:

- XML, XPath
- APIs