

Week 1: Introduction to QTA

POP77142 Quantitative Text Analysis for Social Scientists

Tom Paskhalis

Overview

- Text as data
- Data collection
- Web technologies
- HTML fundamentals
- XPath

Text As Data

Textual Data

- Ubiquitous
- Yet often underutilized

732

22^o—23^o Augusti.

Nathan Butler, Chief Constable, and their Arms taken from them.

The Parties that subscribed it were called in; and did avow the same.

Remitted, upon the Question, That *Nathaniel Butler*, High Constable at *Greenwich*, be forthwith sent for, as a Delinquent: And that *Jo. Rush Painter*, and *Edw. Deneil*, a Vintner, be forthwith summoned to attend the House.

Declaration concerning Constables.

Mr. Reynolds, Mr. Bigley, Mr. Corbett, Sergeant *Wilde*, Sir *Ro. Parkhurst*, Sir *Wm. Arden*, Mr. Solicitor, Mr. *Martin*, Mr. *Giles*, Sir *H. Midway*, Sir *Wm. Ogley*; This Committee is to prepare a Declaration of the Passages this Day informed of, as *Constituted*, that the Kingdom may be informed thereof, that they may put all the strong Towns to make Preparations for their Defence, and to raise the Trained Bands for their Assistance, and to stand upon their Defence: And are to meet this Afternoon, at Four o'Clock, in the Court of Wards.

WHEREAS the King, seduced by traitorous and wicked Counsels, is now in actual War against his Parliament and good People, intending to return an absolute Conqueror of this Kingdom, if not timely prevented; and to this Purpose hath issued out divers illegal Commissions of Array, into all or the most Parts of the Counties of this Kingdom, directed to Lords, and divers other ill-affected Persons, who labour to advance this horrid Design, to the Destruction of the Religion and Liberty of this Kingdom: The Lords and Commons, in Parliament, do therefore *Order, Ordain, and Require*, all Lieutenants, Deputy Lieutenants, Justices of Peace, Sheriffs, Mayors, Bailiffs, Constables, and all other Officers, and all other well-affected Persons whatsoever within this Realm of *England*, and Dominion of *Wales*, to use their utmost Power and Diligence, for the speedy Apprehending, and bringing up to the Parliament, every such Commissioner, and every other Person of what Quality or Condition soever he be, who shall any way countenance or promote the Execution of any such Commission, or that shall raise, levy, muster, or conduct, or endeavour to raise, levy, muster, or conduct any Soldiers or Forces of Horse or Foot, in any Part of this Kingdom, or Dominion of *Wales*, by Virtue or Colour of any such Commission, or other Authority derived from his Majesty, without the Consent of the Lords and Commons: And the Lords and Commons do farther Desire, That all such well-affected Persons as shall employ themselves in perfecting this Ordinance in Execution, shall do an acceptable Service to the State, and shall be defended and protected therein by the Authority of both Houses of Parliament.

Luna, 22^a Augusti, 1642:
Post Meridien.

ORDERED, That the Clerk of the Crown do return the Names, to this House, of the Commissioners of Array in the several and respective Counties.

Remitted, That the Committee of the County of *Cambridgeshire* be enjoined to put the Militia in Execution forthwith, according to the Instructions of the Parliament: And that.

Upon the Report of Mr. *Whitmore*, from the Committee for Informations of divers Priests and others examined before them: It is Ordered, That *Robert Robinson*, a Roman Priest, be committed Prisoner to the King's Bench; *James Browne*, Priest, to the Gatehouse; *Giles Thimbleton* to the New Prison; Dr. *Gibbs* to the Gatehouse.

Ordered, That the Cabinet seized at the Lord Peter's House, be conveyed safely to the Chamber of *London*, that are there to be safely kept, till the House take farther Order.

An Ordinance for the Facts, and Days of Han upon the Question, asse Ordered, The Lords.

Ordered, That the *Wm. Brewster*, be reformed; to consider and Vindication in this the Printing thereof.

Mr. *Cowell* and *Jer* That Mr. *Walton*, a serving the Peace of out his Warrants to a Bands: Mr. *Hegton*, o to be chained up; and the great Diservice of not obey the Parliament rant from the King: V It was Remitted, Th forthwith sent for, as a The Two Witnesses Record in this House, same: Whereupon It was Remitted, up shall be committed t remain during Pleasur Fine, and not be releas Mr. *Coleman* was ca Mr. Speaker propos said Resolution.

Sir *Tho. Barrington* Order for *Colchester* u Fortifications to defen upon.

Ordered, That the from *Manster*, be reu Morning.

Die Martis

PRAYERS.

A LETTER fro 22th of August, Multitude of that To and seizing his Hors and that they would a and his Lady were co Newcomen, Clerk, wa to the King; and th should have gone, an Remitted, That Sir, Colchester, Clerk, sh in safe Custody, by th Remitted, upon the Ammunition, Monies, his House (a perfect Inventory being first made), shall be brought up to London: That those Gentlemen of Essex that are appointed to be sent to Colchester, shall have Power to send up such Persons in safe Custody, as, upon Examination, they shall think necessary.

Sir *Tho. Barrington*, Mr. *Grimston* are appointed to go to Colchester to apprese the Tumult there.

Mr. *Reynolds*, Mr. *Martin*, Sir *Tho. Barrington*, Mr. *Prisdeman*, are appointed to prepare a Declaration to be sent to Colchester, and Instructions to accompany those Gentlemen that are appointed to go to Colchester.

Mr. *Spartaco*, Mr. *Bence*, Mr. *Fenne*, are appointed to remove repair to Sir *Jo. Heyden*'s House; and to view the Arms, Ammunition, and other warlike Provisions, that are there; and to remove them from thence; and to lay them up in some safe Place in the City.

Ordered,

Τα σόσιαλ μίντια καταστρέφουν τη Δημοκρατία

Η δημοσιογράφος, συγγραφέας και εδώ και λίγες μέρες κάτοχος του Νόμπελ Ειρήνης Μαρία Ρέσα μιλά στην Καθημερινή



Newsroom
24.10.2021 • 12:10

Η δημοσιογράφος, συγγραφέας και εδώ και λίγες μέρες κάτοχος του Νόμπελ Ειρήνης Μαρία Ρέσα μίλησε στην Καθημερινή για το πικρό μήνυμα που στέλνει η βράβευσή της και του Ρώσου δημοσιογράφου Ντμίτρι Μουράτοφ σε μια εποχή που η Δημοκρατία απειλείται από την άνοδο ακραίων δυνάμεων και τη διάδοση των fake news μέσω των σόσιαλ μίντια.

Web Scraping

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, Playwright - Python bindings recommended)
- Web API: sending requests and processing responses (HTTP queries, `httr2` in R, `requests` in Python)

Web Tables



Article [Talk](#)

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

(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=1

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

 **EUR-Lex**
Access to European Union law

English  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

   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 [\(177366\)](#)

Search Results

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

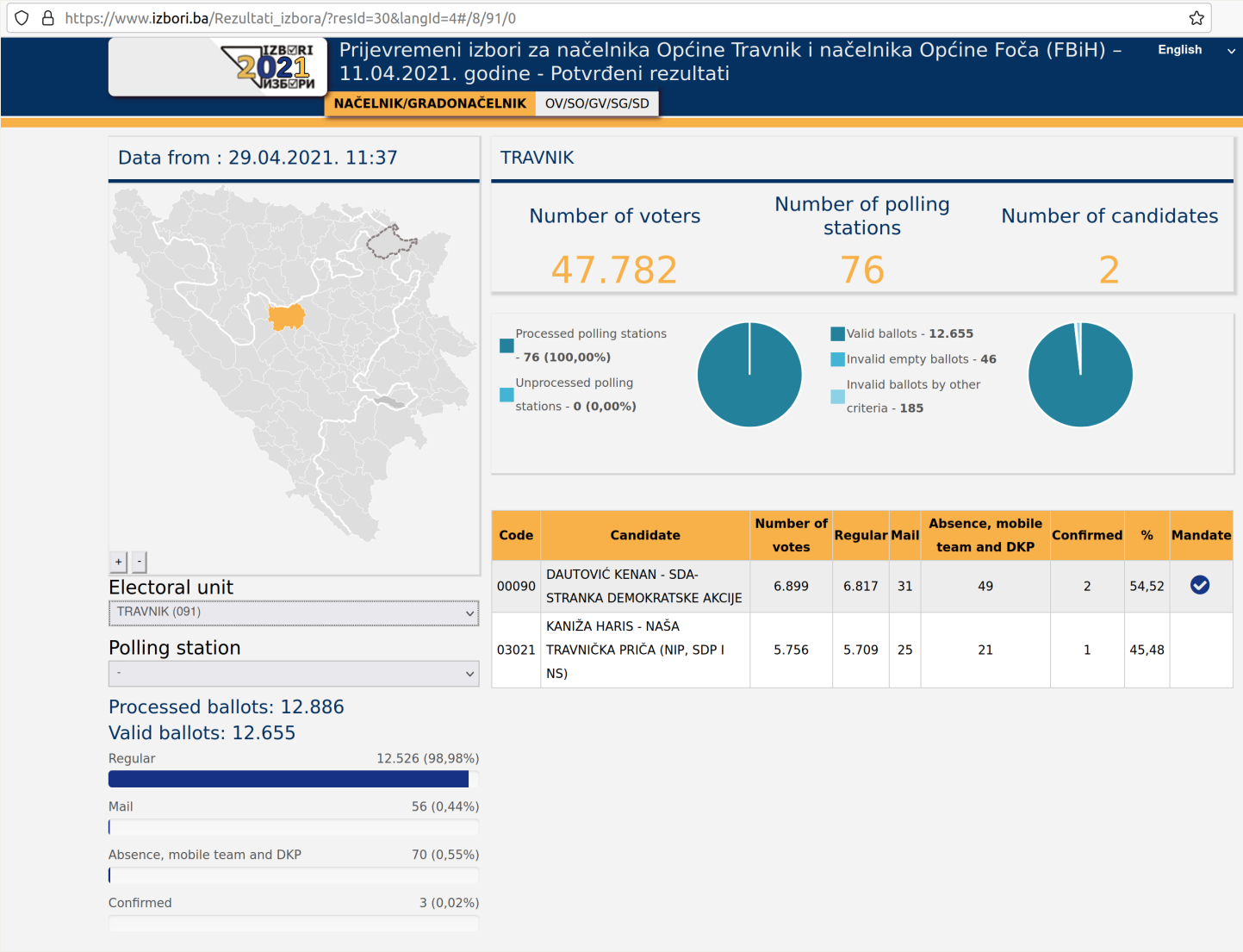
> Search criteria

☐ Results 1 - 10 of 1021911 [Sort by](#) Relevant  

[1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [11](#) [12](#) [13](#) [14](#) [15](#) [16](#) [17](#) [18](#) [19](#) [20](#) [21](#) [22](#) [23](#) [24](#) [25](#) [26](#) [27](#) [28](#) [29](#) [30](#) [31](#) [32](#) [33](#) [34](#) [35](#) [36](#) [37](#) [38](#) [39](#) [40](#) [41](#) [42](#) [43](#) [44](#) [45](#) [46](#) [47](#) [48](#) [49](#) [50](#) [51](#) [52](#) [53](#) [54](#) [55](#) [56](#) [57](#) [58](#) [59](#) [60](#) [61](#) [62](#) [63](#) [64](#) [65](#) [66](#) [67](#) [68](#) [69](#) [70](#) [71](#) [72](#) [73](#) [74](#) [75](#) [76](#) [77](#) [78](#) [79](#) [80](#) [81](#) [82](#) [83](#) [84](#) [85](#) [86](#) [87](#) [88](#) [89](#) [90](#) [91](#) [92](#) [93](#) [94](#) [95](#) [96](#) [97](#) [98](#) [99](#) [100](#) [101](#) [102](#) [103](#) [104](#) [105](#) [106](#) [107](#) [108](#) [109](#) [110](#) [111](#) [112](#) [113](#) [114](#) [115](#) [116](#) [117](#) [118](#) [119](#) [120](#) [121](#) [122](#) [123](#) [124](#) [125](#) [126](#) [127](#) [128](#) [129](#) [130](#) [131](#) [132](#) [133](#) [134](#) [135](#) [136](#) [137](#) [138](#) [139](#) [140](#) [141](#) [142](#) [143](#) [144](#) [145](#) [146](#) [147](#) [148](#) [149](#) [150](#) [151](#) [152](#) [153](#) [154](#) [155](#) [156](#) [157](#) [158](#) [159](#) [160](#) [161](#) [162](#) [163](#) [164](#) [165](#) [166](#) [167](#) [168](#) [169](#) [170](#) [171](#) [172](#) [173](#) [174](#) [175](#) [176](#) [177](#) [178](#) [179](#) [180](#) [181](#) [182](#) [183](#) [184](#) [185](#) [186](#) [187](#) [188](#) [189](#) [190](#) [191](#) [192](#) [193](#) [194](#) [195](#) [196](#) [197](#) [198](#) [199](#) [200](#) [201](#) [202](#) [203](#) [204](#) [205](#) [206](#) [207](#) [208](#) [209](#) [210](#) [211](#) [212](#) [213](#) [214](#) [215](#) [216](#) [217](#) [218](#) [219](#) [220](#) [221](#) [222](#) [223](#) [224](#) [225](#) [226](#) [227](#) [228](#) [229](#) [230](#) [231](#) [232](#) [233](#) [234](#) [235](#) [236](#) [237](#) [238](#) [239](#) [240](#) [241](#) [242](#) [243](#) [244](#) [245](#) [246](#) [247](#) [248](#) [249](#) [250](#) [251](#) [252](#) [253](#) [254](#) [255](#) [256](#) [257](#) [258](#) [259](#) [260](#) [261](#) [262](#) [263](#) [264](#) [265](#) [266](#) [267](#) [268](#) [269](#) [270](#) [271](#) [272](#) [273](#) [274](#) [275](#) [276](#) [277](#) [278](#) [279](#) [280](#) [281](#) [282](#) [283](#) [284](#) [285](#) [286](#) [287](#) [288](#) [289](#) [290](#) [291](#) [292](#) [293](#) [294](#) [295](#) [296](#) [297](#) [298](#) [299](#) [300](#) [301](#) [302](#) [303](#) [304](#) [30](#)

(Eur-Lex)

Interactive Webpages



(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

Web Technologies

- Key technologies used to disseminate content on the Web:
 - XML/HTML (Extensible Markup Language/Hypertext Markup Language)
 - CSS (Cascading Style Sheets)
 - JavaScript
 - API (Application Programming Interface)
 - JSON (JavaScript Object Notation)

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
 - They can contain elements that change the appearance of a website
 - Example: POP77142 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**ypertext **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

```
<!DOCTYPE html>
<html>
  <head>
    <title>A title</title>
  </head>
  <body>
    <h1 style="color:Red;">A heading</h1>
    <p>A paragraph.</p>
  </body>
</html>
```

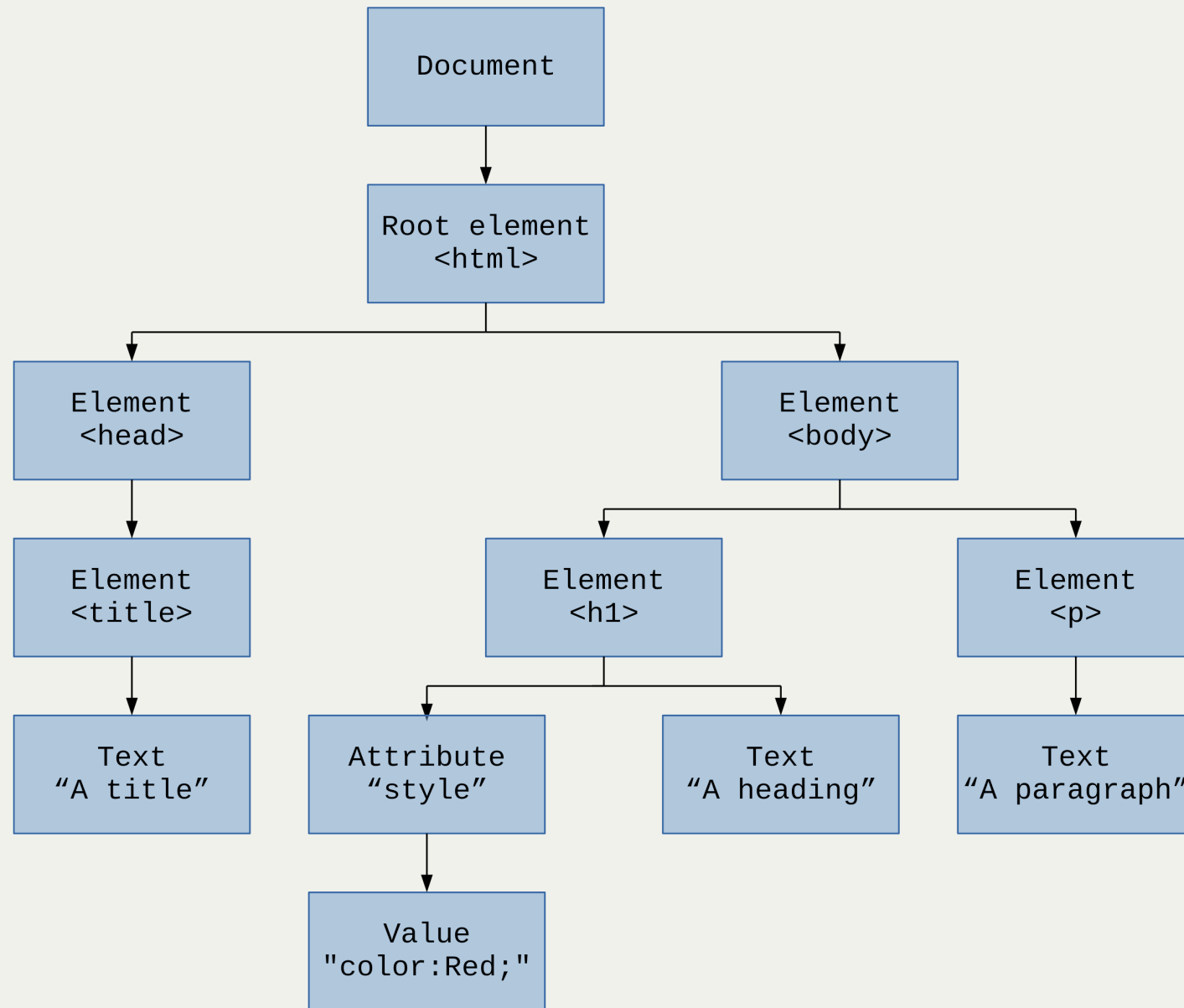


W3Schools: Try HTML

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 an element is found between the 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>`

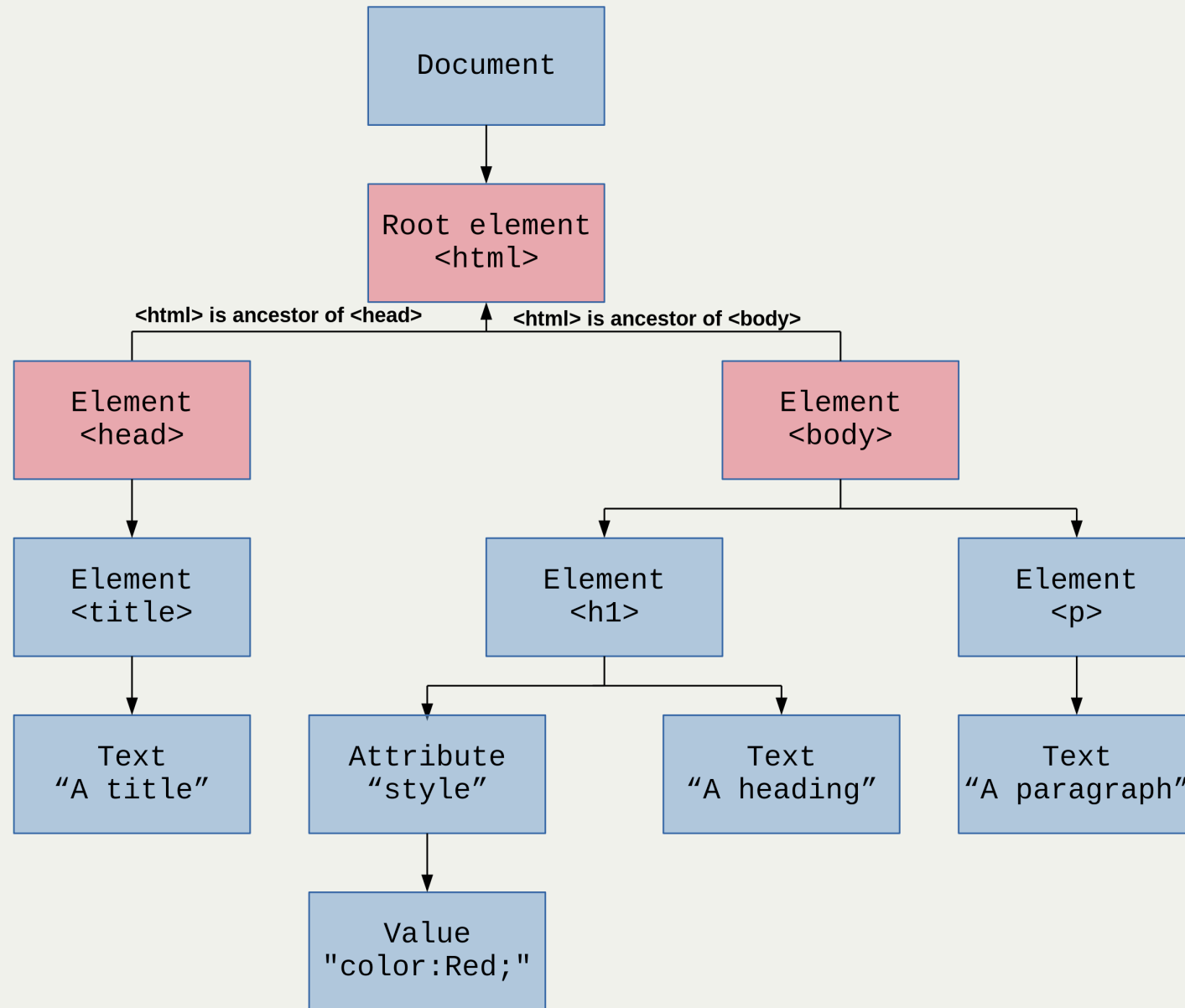
HTML tree



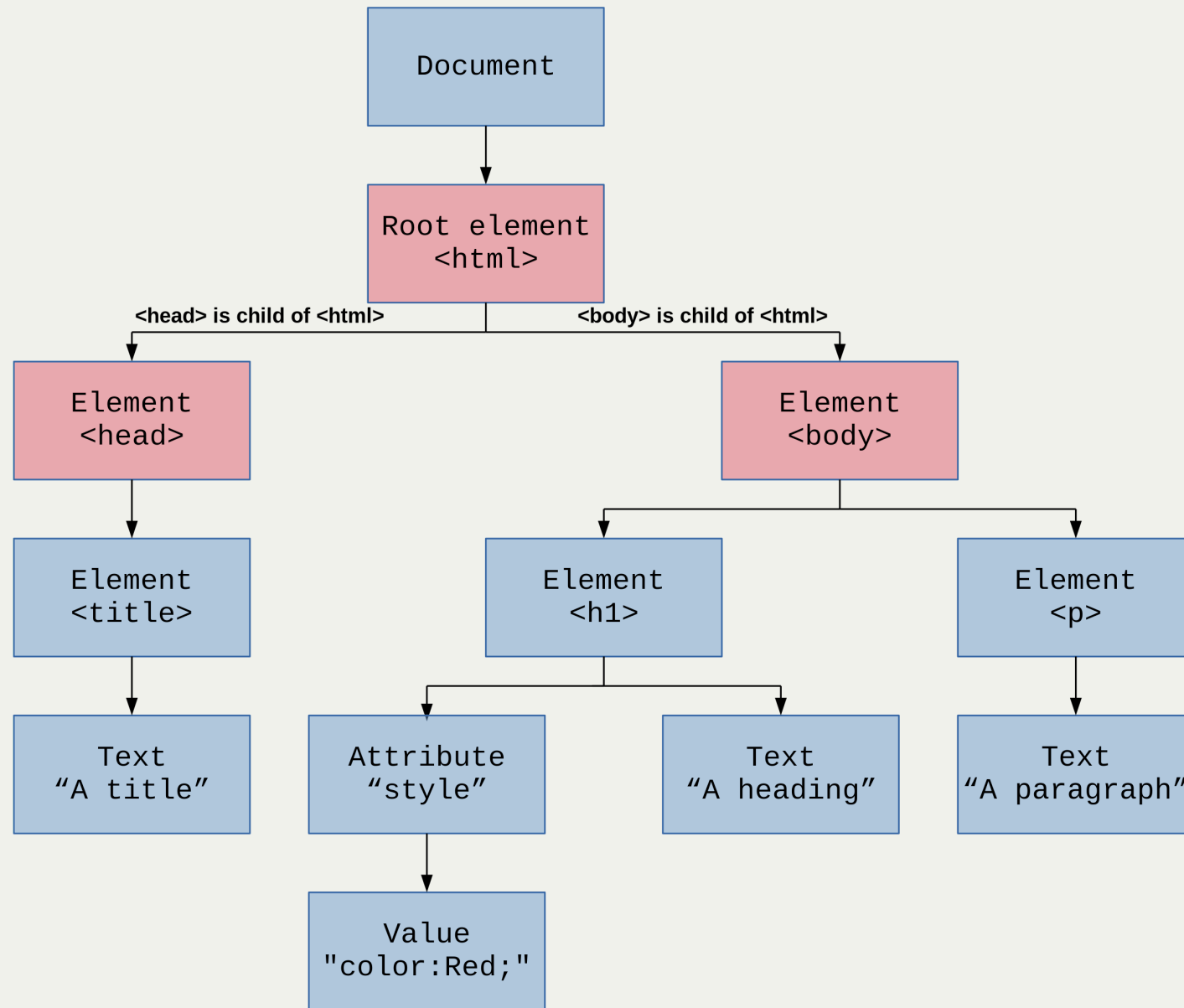
HTML Tree Relationships

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

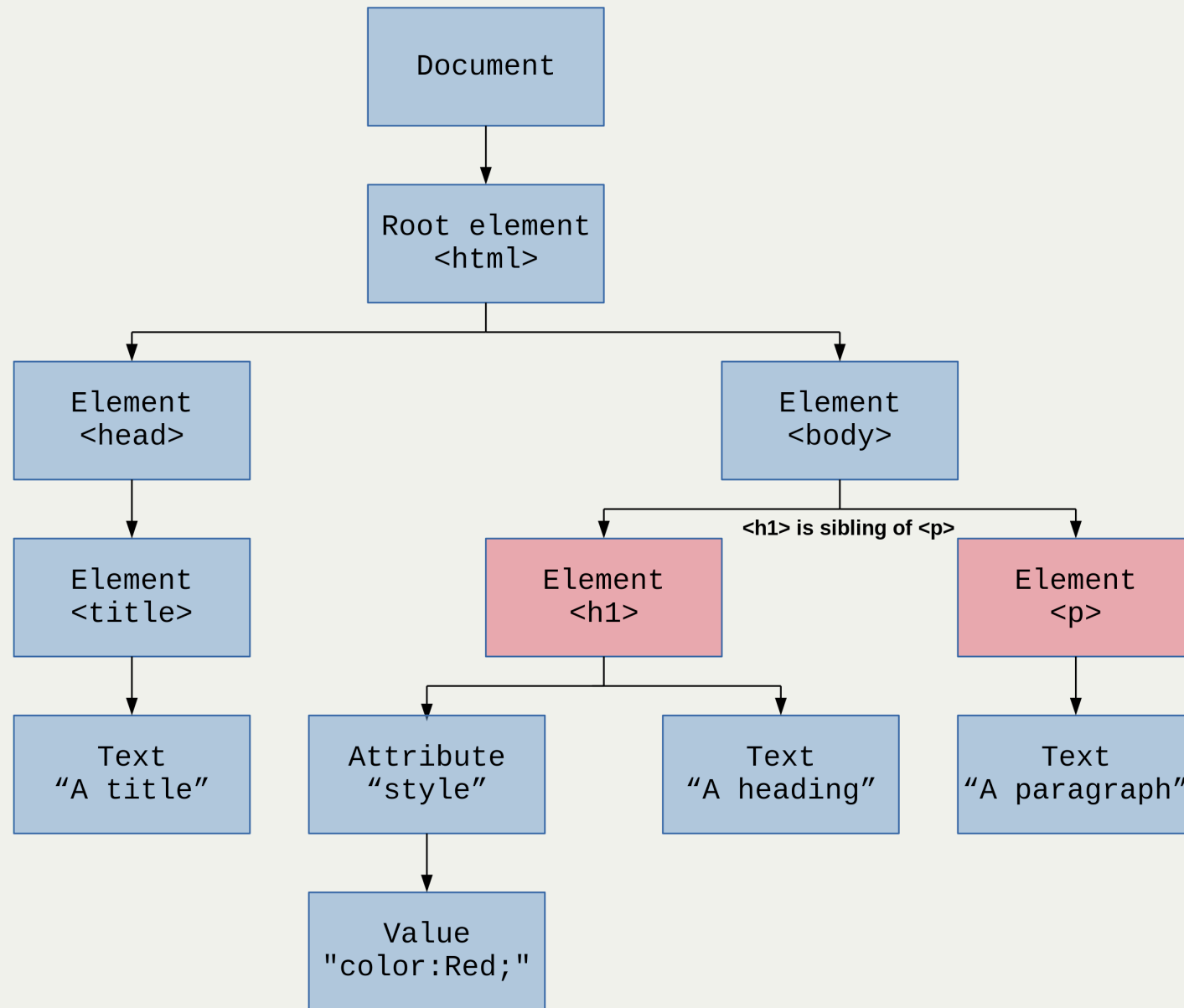
HTML Parent/Ancestor



HTML Children/Descendants



HTML Siblings



Parsing HTML Tree: Example

```
1 library("rvest")
```

```
1 html_txt <- "  
2 <!DOCTYPE html>  
3 <html>  
4   <head>  
5     <title>A title</title>  
6   </head>  
7   <body>  
8     <h1 style='color:Red;'>A heading</h1>  
9     <p>A paragraph.</p>  
10  </body>  
11 </html>"
```

```
1 html <- rvest::read_html(html_txt)
```

```
1 str(html)
```

List of 2

\$ node:<externalptr>

\$ doc :<externalptr>

- attr(*, "class")= chr [1:2] "xml_document" "xml_node"

Parsing HTML Tree: Example

```
1 children <- rvest::html_children(html)
2 children
```

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

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

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

```
1 children2 <- rvest::html_children(body)
2 children2
```

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

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

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

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

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

XML: Extensible Markup Language

- XML (Extensible Markup Language) is a more general form of markup language
- Allows sharing structured data of tree-like form
- Relative to HTML:
 - Tags are user-defined
 - End tags are always required
 - Stricter (no inconsistencies permitted)

```
<?xml version="1.0" encoding="UTF-8" ?>
<courses>
  <course>
    <title>Computer Programming for Social Scientists</title>
    <code>POP77001</code>
    <year>2024</year>
    <term>Michaelmas</term>
    <description>Course on computer programming in Python and R.
  </description>
  </course>
  <course>
    <title>Quantitative Text Analysis for Social
Scientists</title>
    <code>POP77142</code>
    <year>2025</year>
    <term>Hillary</term>
    <description>Introduction to text analysis.</description>
  </course>
</courses>
```

Parsing XML Tree: Example

```
1 library("xml2")

1 xml_txt <-
2 '<?xml version="1.0" encoding="UTF-8" ?>
3 <courses>
4   <course>
5     <title>Computer Programming for Social Scientists</title>
6     <code>POP77001</code>
7     <year>2024</year>
8     <term>Michaelmas</term>
9     <description>Course on computer programming in Python and R.</description>
10  </course>
11  <course>
12    <title>Quantitative Text Analysis for Social Scientists</title>
13    <code>POP77142</code>
14    <year>2025</year>
15    <term>Hillary</term>
16    <description>Introduction to text analysis.</description>
17  </course>
18 </courses>'

1 xml <- xml2::read_xml(xml_txt)

1 str(xml)
```

List of 2

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

Parsing XML Tree: Example

```
1 children3 <- xml2::xml_children(xml)
2 children3
```

```
{xml_nodeset (2)}
[1] <course>\n  <title>Computer Programming for Social Scientists</title>\n  ...
[2] <course>\n  <title>Quantitative Text Analysis for Social Scientists</titl ...
```

```
1 pop77001 <- children3[1]
2 xml2::xml_children(pop77001)
```

```
{xml_nodeset (5)}
[1] <title>Computer Programming for Social Scientists</title>
[2] <code>POP77001</code>
[3] <year>2024</year>
[4] <term>Michaelmas</term>
[5] <description>Course on computer programming in Python and R.</description>
```

```
1 xml2::xml_text(xml_children(children3[1]))
```

```
[1] "Computer Programming for Social Scientists"
[2] "POP77001"
[3] "2024"
[4] "Michaelmas"
[5] "Course on computer programming in Python and R."
```

Examples of XML

- RSS (**R**eally **S**imple **S**yndication) feeds
- SVG (**S**calable **V**ector **G**raphics) images
- Modern office documents (Microsoft Office `.docx`, `.xlsx`, `.pptx`, OpenOffice/LibreOffice)

Parsing XML/HTML with XPath

- XPath (XML Path Language) is a language for selecting parts of XML/HTML tree
- Basic syntax:
 - `/` - select element at the root node (e.g. `/html/body`)
 - `//` - select element at any depth (e.g. `//h1`)
 - `//<tag>/*` - select all descendants of tag (e.g. `//body/*`)
 - `//<tag>[@<attr>]` - select all elements that have given attribute (e.g. `//h1[@style]`)
 - `//<tag>[@<attr>='<value>']` - select all elements, whose attribute has given value (e.g. `//h1[@style='color:Red;']`)



Extra

XPath syntax

Parsing XML/HTML with XPath: Example

```
1 rvest::html_elements(html, xpath = "//p")
```

```
{xml_nodeset (1)}  
[1] <p>A paragraph.</p>
```

```
1 rvest::html_elements(html, xpath = "//h1[@style='color:Red;']")
```

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

```
1 xml2::xml_find_all(xml, xpath = "//code")
```

```
{xml_nodeset (2)}  
[1] <code>POP77001</code>  
[2] <code>POP77142</code>
```

```
1 # We can also find elements by text  
2 xml2::xml_find_all(xml, xpath = "//code[text()='POP77001']")
```

```
{xml_nodeset (1)}  
[1] <code>POP77001</code>
```

Scraping Webpage



Article [Talk](#)

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

(Wikipedia)

Scraping Webpage with XPath: Example

```
1 html <- rvest::read_html("https://en.wikipedia.org/wiki/Members_of_the_1st_D%C3%A1il")
```

```
1 tables <- rvest::html_elements(html, xpath = "//table")
2 tables
```

```
{xml_nodeset (8)}
[1] <table class="box-More_citations_needed plainlinks metadata ambox ambox-c ...
[2] <table class="infobox vevent"><tbody>\n<tr><th colspan="2" class="infobox ...
[3] <table style="width:100%; border-collapse:collapse"><tbody><tr style="ver ...
[4] <table class="wikitable" style="font-size: 95%;"><tbody>\n<tr>\n<th colsp ...
[5] <table class="wikitable sortable"><tbody>\n<tr>\n<th rowspan="2">Constitu ...
[6] <table class="wikitable"><tbody>\n<tr>\n<th>Constituency\n</th>\n<th>Outg ...
[7] <table class="wikitable"><tbody>\n<tr>\n<th>Winner\n</th>\n<th colspan="2 ...
[8] <table class="nowraplinks mw-collapsible autocollapse navbox-inner" style ...
```

```
1 tbody <- rvest::html_children(tables[5])
2 tbody
```

```
{xml_nodeset (1)}
[1] <tbody>\n<tr>\n<th rowspan="2">Constituency\n</th>\n<th rowspan="2">Name\ ...
```

```
1 tds <- rvest::html_table(tbody)
2 tds
```

```
[[1]]
# A tibble: 109 × 8
  Constituency Name Portrait `Party affiliation` `Party affiliation`
  <chr>         <chr>      <chr>      <chr>          <chr>
1 Constituency Name "Portra... "Start of Dáil ter... Start of Dáil term
2 Antrim East Robert McCal... "" "" Irish Unionist
```

				"Elected in 1919 b...	Elected in 1919 by...
3	Antrim East	George Hanna	""	""	
4	Antrim Mid	Hugh O'Neill	""	""	Irish Unionist
5	Antrim North	Peter Kerr-S...	""	""	Irish Unionist
6	Antrim South	Charles Curt...	""	""	Irish Unionist
7	Armagh Mid	James Rolsto...	""	""	Irish Unionist
8	Armagh North	William Allen	""	""	Irish Unionist
9	Armagh South	Patrick Donn...	""	""	Irish Parliamentary
10	Belfast Cromac	William Arth...	""	""	Irish Unionist

i 99 more rows

i 3 more variables: `Party affiliation` <chr>, `Party affiliation` <chr>,
`Assumed office` <chr>

Scraping Webpage with XPath:

Example

```
1 str(tds)
```

List of 1

```
$ : tibble [109 × 8] (S3: tbl_df/tbl/data.frame)
  ..$ Constituency      : chr [1:109] "Constituency" "Antrim East" "Antrim East" "Antrim Mid" ...
  ..$ Name              : chr [1:109] "Name" "Robert McCalmont" "George Hanna" "Hugh O'Neill" ...
  ..$ Portrait          : chr [1:109] "Portrait" "" "" "" ...
  ..$ Party affiliation: chr [1:109] "Start of Dáil term" "" "Elected in 1919 by-electionas Independent
Unionist" "" ...
  ..$ Party affiliation: chr [1:109] "Start of Dáil term" "Irish Unionist" "Elected in 1919 by-electionas
Independent Unionist" "Irish Unionist" ...
  ..$ Party affiliation: chr [1:109] "End of Dáil term" "Resigned in 1919" "" "" ...
  ..$ Party affiliation: chr [1:109] "End of Dáil term" "Resigned in 1919" "Ulster Unionist" "Ulster Unionist"
...
  ..$ Assumed office    : chr [1:109] "Assumed office" "Abstained" "Abstained" "Abstained" ...
```

```
1 tds <- tds[[1]]
2 head(tds)
```

```
# A tibble: 6 × 8
  Constituency Name      Portrait `Party affiliation` `Party affiliation`
  <chr>             <chr>      <chr>      <chr>             <chr>
1 Constituency Name    "Portra... "Start of Dáil ter... Start of Dáil term
2 Antrim East  Robert McCalmont ""          ""              Irish Unionist
3 Antrim East  George Hanna   ""          "Elected in 1919 b... Elected in 1919 by...
4 Antrim Mid   Hugh O'Neill   ""          ""              Irish Unionist
5 Antrim North Peter Kerr-Smil... ""          ""              Irish Unionist
6 Antrim South Charles Curtis ... ""          ""              Irish Unionist
```

Scraping Webpage with XPath: Example

```
1 colnames(tds) <- tds[1,]
2 tds <- tds[-1,]
3 head(tds)
```

```
# A tibble: 6 × 8
```

	Constituency	Name	Portrait	`Start of Dáil term`	`Start of Dáil term`
	<chr>	<chr>	<chr>	<chr>	<chr>
1	Antrim East	Robert McCalm...	""	""	Irish Unionist
2	Antrim East	George Hanna	""	"Elected in 1919 by...	Elected in 1919 by-...
3	Antrim Mid	Hugh O'Neill	""	""	Irish Unionist
4	Antrim North	Peter Kerr-Sm...	""	""	Irish Unionist
5	Antrim South	Charles Curti...	""	""	Irish Unionist
6	Armagh Mid	James Rolston...	""	""	Irish Unionist

```
# i 3 more variables: `End of Dáil term` <chr>, `End of Dáil term` <chr>,
#   `Assumed office` <chr>
```

```
1 tds <- tds[, -3]
2 str(tds)
```

```
tibble [108 × 7] (S3: tbl_df/tbl/data.frame)
```

```
$ Constituency      : chr [1:108] "Antrim East" "Antrim East" "Antrim Mid" "Antrim North" ...
$ Name              : chr [1:108] "Robert McCalmont" "George Hanna" "Hugh O'Neill" "Peter Kerr-Smiley" ...
$ Start of Dáil term: chr [1:108] "" "Elected in 1919 by-electionas Independent Unionist" "" "" ...
$ Start of Dáil term: chr [1:108] "Irish Unionist" "Elected in 1919 by-electionas Independent Unionist" "Irish
Unionist" "Irish Unionist" ...
$ End of Dáil term  : chr [1:108] "Resigned in 1919" "" "" "" ...
$ End of Dáil term  : chr [1:108] "Resigned in 1919" "Ulster Unionist" "Ulster Unionist" "Ulster Unionist" ...
$ Assumed office    : chr [1:108] "Abstained" "Abstained" "Abstained" "Abstained" ...
```

Web Scraping in Practice

- Always check first whether an API for querying exists.
- It is the most robust (and sanctioned) way of obtaining data.
- Check copyrights and respect those when using scraped data.
- Limit you scraping bandwidth (introduce waiting times between queries).

Next

- Tutorial: HTML and web scraping
- Assignment 1: Due 15:59 on Wednesday, 26th March
(submission on Blackboard)