Week 1: Introduction to QTA

POP77032 Quantitative Text Analysis for Social Scientists

Tom Paskhalis

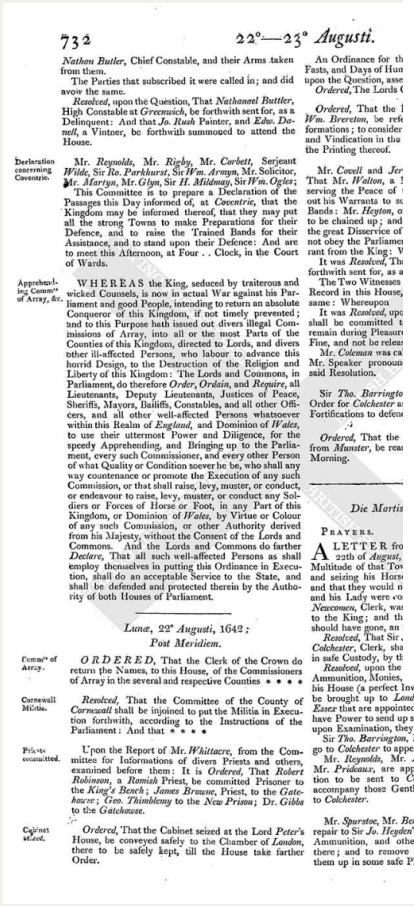
Overview

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

Text As Data

Textual Data

- Ubiquitous
- Yet often underutilized



732 22°—23° *Augusti.*

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

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

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

Declaration concerning Coventrie.

Mr. *Reynolds*, Mr. *Rigby*, Mr. *Corbett*, Serjeant *Wilde*, Sir *Ro. Parkhurst*, Sir *Wm. Armyn*, Mr. Solicitor, Mr. *Martyn*, Mr. *Glyn*, Sir *H. Mildmay*, Sir *Wm. Ogle*;

This Committee is to prepare a Declaration of the Passages this Day informed of, at *Coventrie*, 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 . . Clock, in the Court of Wards.

Apprehending Comm^rs of Array, &c.

WHEREAS the King, seduced by traiterous 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 uttermost 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 *Declare*, That all such well-affected Persons as shall employ themselves in putting 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° *Augusti*, 1642;
Post Meridiem.

Comm^rs of Array.

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 * * * *

Cornwall Militia.

Resolved, That the Committee of the County of *Cornwall* shall be injoined to put the Militia in Execution forthwith, according to the Instructions of the Parliament: And that * * * *

Priests committed.

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

Cabinet seized.

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

An Ordinance for th
Fasts, and Days of Hum
upon the Question, asse
Ordered, The Lords

Ordered, That the
Wm. Brereton, be ref
formations; to consider
and Vindication in tha
the Printing thereof.

Mr. *Cowell* and *Jer*
That Mr. *Walton*, a
serving the Peace of
out his Warrants to s
Bands: Mr. *Heyton*, o
to be chained up; and
the great Disservice of
not obey the Parliamen
rant from the King: W
It was *Resolved*, Th
forthwith sent for, as a
The Two Witnesses
Record in this House,
same: Whereupon
It was *Resolved*, upo
shall be committed t
remain during Pleasur
Fine, and not be relea
Mr. *Coleman* was ca
Mr. Speaker pronoun
said Resolution.

Sir *Tho. Barrington*
Order for *Colchester* a
Fortifications to defen

Ordered, That the
from *Munster*, be rea
Morning.

Die Martis

PRAYERS.

A LETTER fro
22th of *August*,
Multitude of that To
and seizing his Hors
and that they would n
and his Lady were co
Newcomen, Clerk, wa
to the King; and th
should have gone, an
Resolved, That Sir
Colchester, Clerk, sha
in safe Custody, by th
Resolved, 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 appease the Tumults there.

Mr. *Reynolds*, Mr. *Marten*, Sir *Tho. Barrington*, Mr. *Prideaux*, 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. *Spurstoe*, Mr. *Bence*, Mr. *Venne*, are appointed to 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.

Removing Arms.

Ordered,

Elon Musk @elonmusk · 13 May

Twitter deal temporarily on hold pending details supporting calculation that spam/fake accounts do indeed represent less than 5% of users

reuters.c
Twitter es
Twitter In
represent

26.2K

Elon Musk
@elonmusk

To find out, m
followers of @

I invite others
they discover

2:47 am · 14 May 20

10.4K Retweets

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 ηχηρό μήνυμα που στέλνει η βράβευση 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

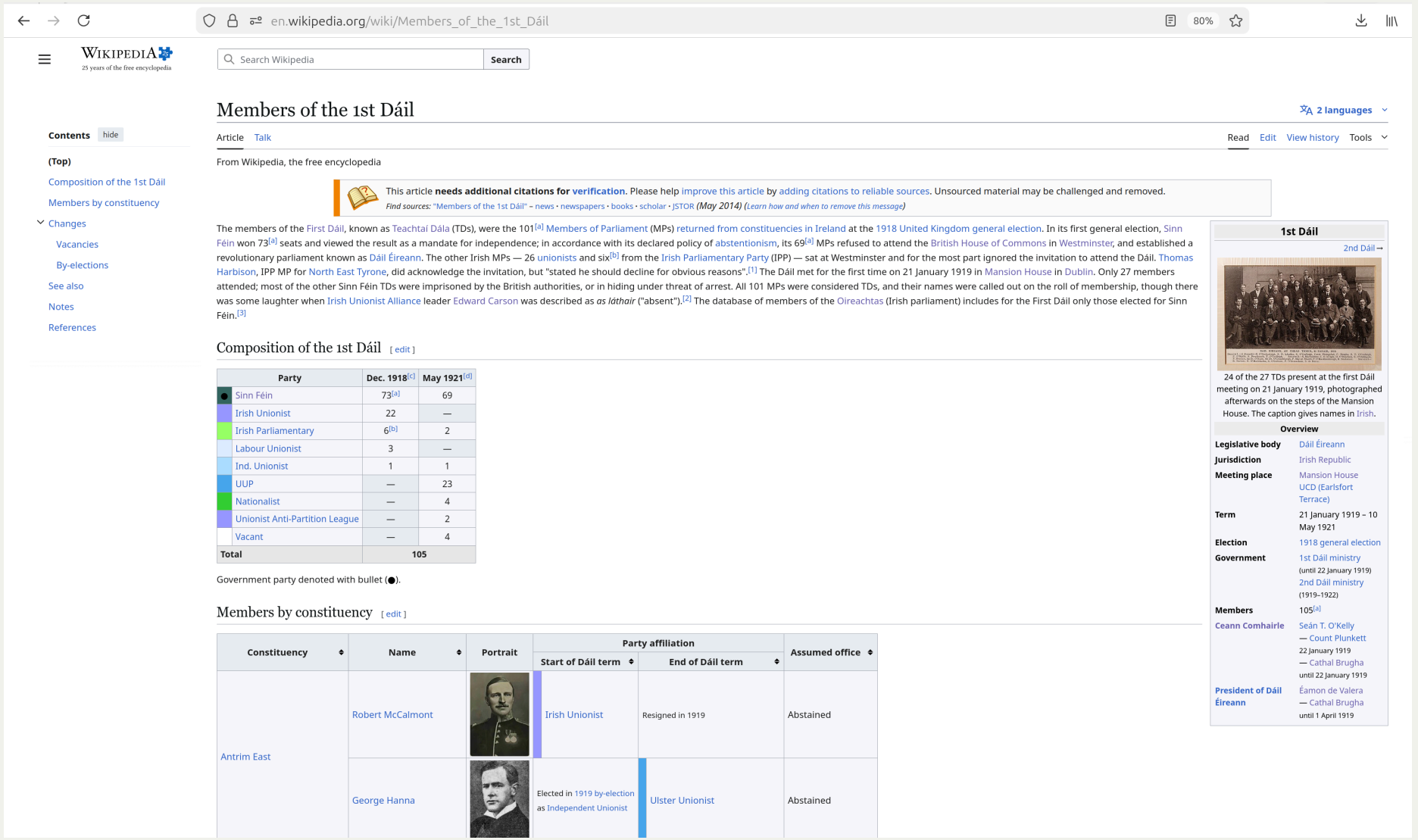
Online Data Sources

- 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



en.wikipedia.org/wiki/Members_of_the_1st_Dáil

Wikipedia
25 years of the free encyclopedia

Search Wikipedia | Search

Members of the 1st Dáil

2 languages

Article | Talk | Read | Edit | View history | Tools

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "Members of the 1st Dáil" – news · newspapers · books · scholar · JSTOR *(May 2014)* (*Learn how and when to remove this message*)

Contents hide

(Top)
Composition of the 1st Dáil
Members by constituency
Changes
Vacancies
By-elections
See also
Notes
References

The members of the First Dáil, known as Teachtaí Dála (TDs), were the 101[a] Members of Parliament (MPs) returned from constituencies in Ireland at the 1918 United Kingdom general election. In its first general election, Sinn Féin won 73[a] seats and viewed the result as a mandate for independence; in accordance with its declared policy of abstentionism, its 69[a] MPs refused to attend the British House of Commons in Westminster, and established a revolutionary parliament known as Dáil Éireann. The other Irish MPs — 26 unionists and six[b] from the Irish Parliamentary Party (IPP) — sat at Westminster and for the most part ignored the invitation to attend the Dáil. Thomas Harbison, IPP MP for North East Tyrone, did acknowledge the invitation, but "stated he should decline for obvious reasons".[1] The Dáil met for the first time on 21 January 1919 in Mansion House in Dublin. Only 27 members attended; most of the other Sinn Féin TDs were imprisoned by the British authorities, or in hiding under threat of arrest. All 101 MPs were considered TDs, and their names were called out on the roll of membership, though there was some laughter when Irish Unionist Alliance leader Edward Carson was described as *as láthair* ("absent").[2] The database of members of the Oireachtas (Irish parliament) includes for the First Dáil only those elected for Sinn Féin.[3]

1st Dáil
2nd Dáil →

24 of the 27 TDs present at the first Dáil meeting on 21 January 1919, photographed afterwards on the steps of the Mansion House. The caption gives names in Irish.

Overview	
Legislative body	Dáil Éireann
Jurisdiction	Irish Republic
Meeting place	Mansion House UCD (Earlsfort Terrace)
Term	21 January 1919 – 10 May 1921
Election	1918 general election
Government	1st Dáil ministry (until 22 January 1919) 2nd Dáil ministry (1919–1922)
Members	105[a]
Ceann Comhairle	Seán T. O'Kelly — Count Plunkett 22 January 1919 — Cathal Brugha until 22 January 1919
President of Dáil Éireann	Éamon de Valera — Cathal Brugha until 1 April 1919

Composition of the 1st Dáil [edit]

Party	Dec. 1918[c]	May 1921[d]
● Sinn Féin	73[a]	69
Irish Unionist	22	—
Irish Parliamentary	6[b]	2
Labour Unionist	3	—
Ind. Unionist	1	1
UUP	—	23
Nationalist	—	4
Unionist Anti-Partition League	—	2
Vacant	—	4
Total	**105**	

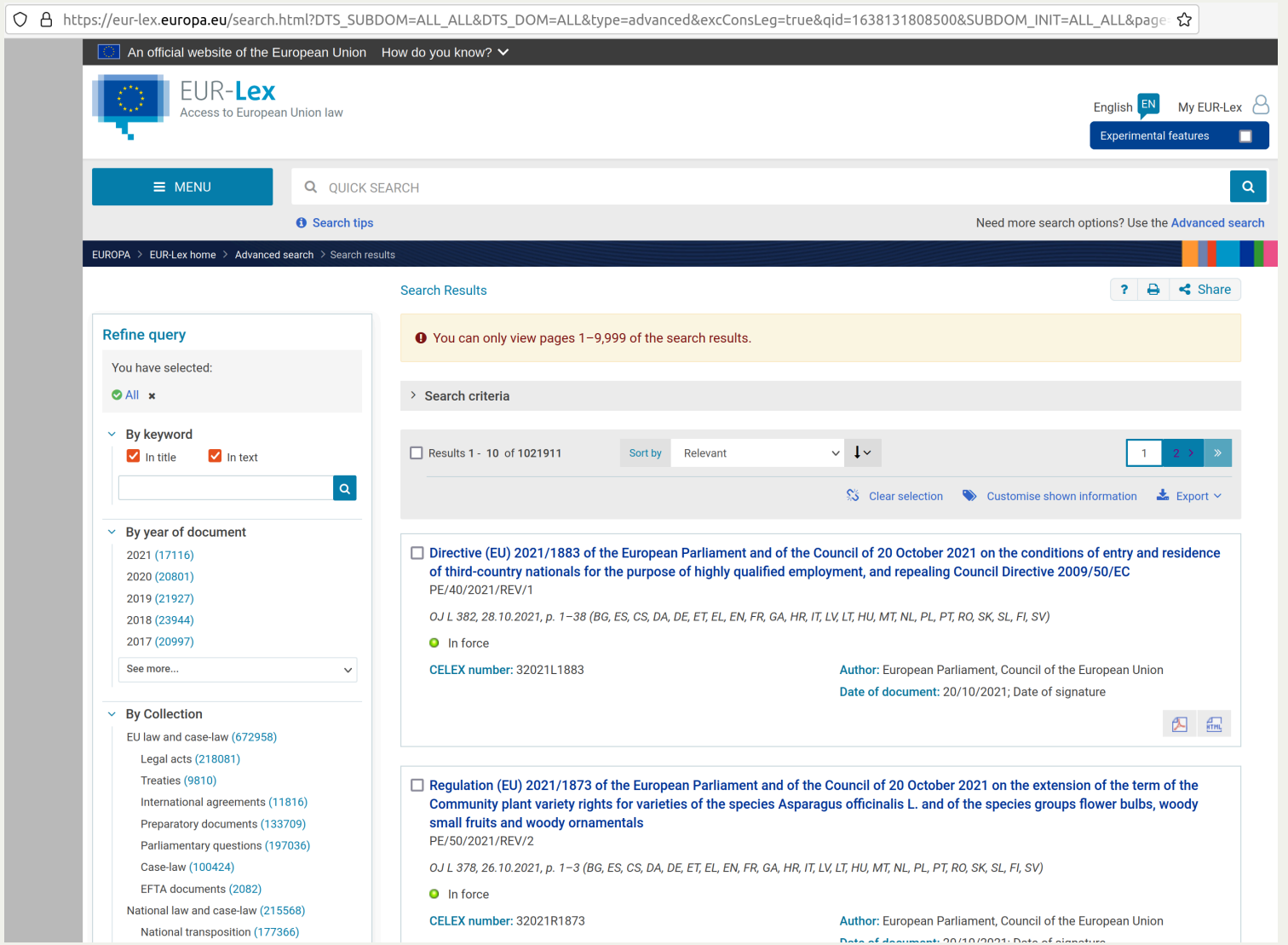
Government party denoted with bullet (●).

Members by constituency [edit]

Constituency	Name	Portrait	Party affiliation: Start of Dáil term	Party affiliation: End of Dáil term	Assumed office
Antrim East	Robert McCalmont		Irish Unionist	Resigned in 1919	Abstained
	George Hanna		Elected in 1919 by-election as Independent Unionist	Ulster Unionist	Abstained

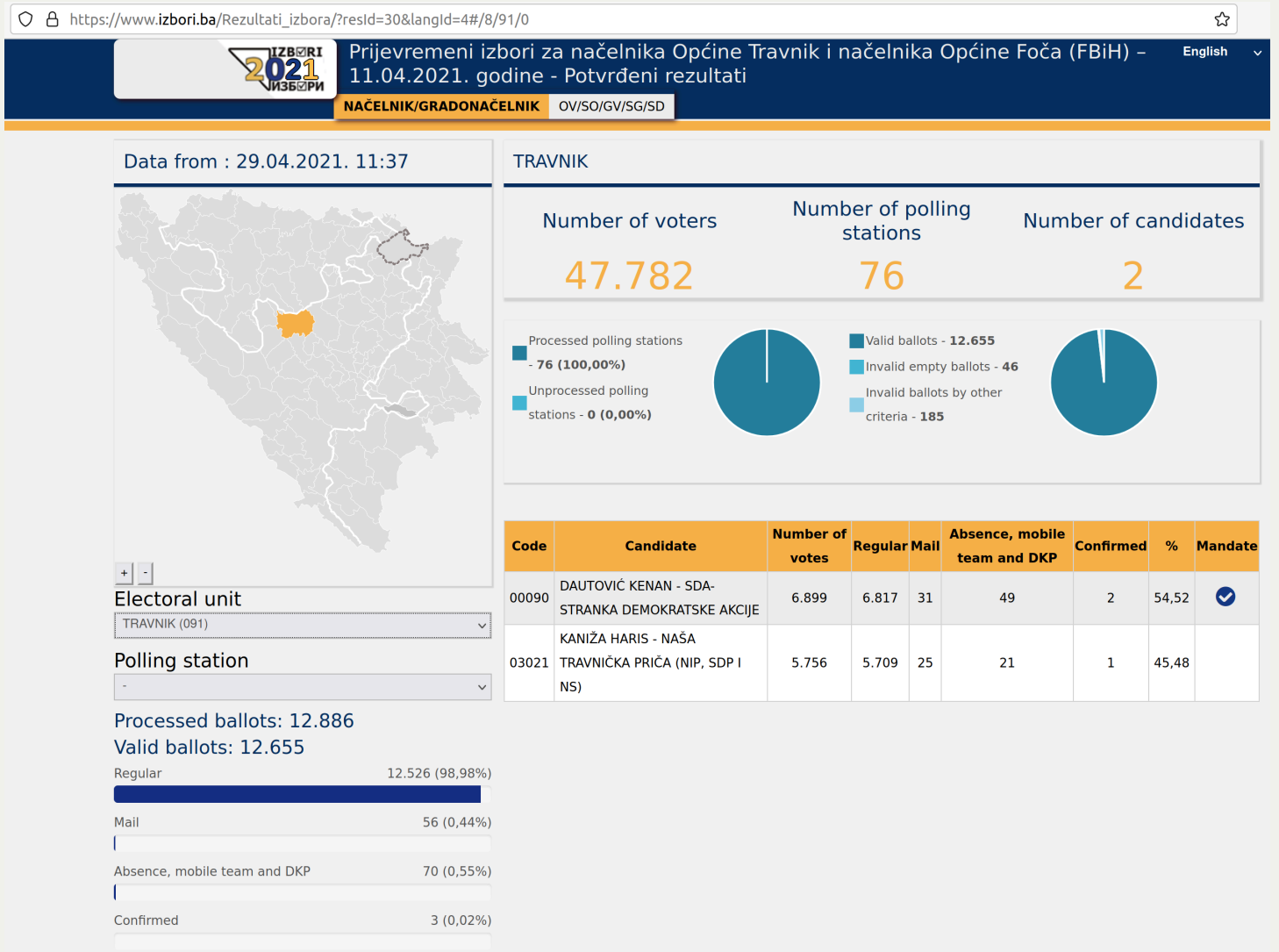
(Wikipedia)

Unstructured Data



(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 (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
  - 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**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

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

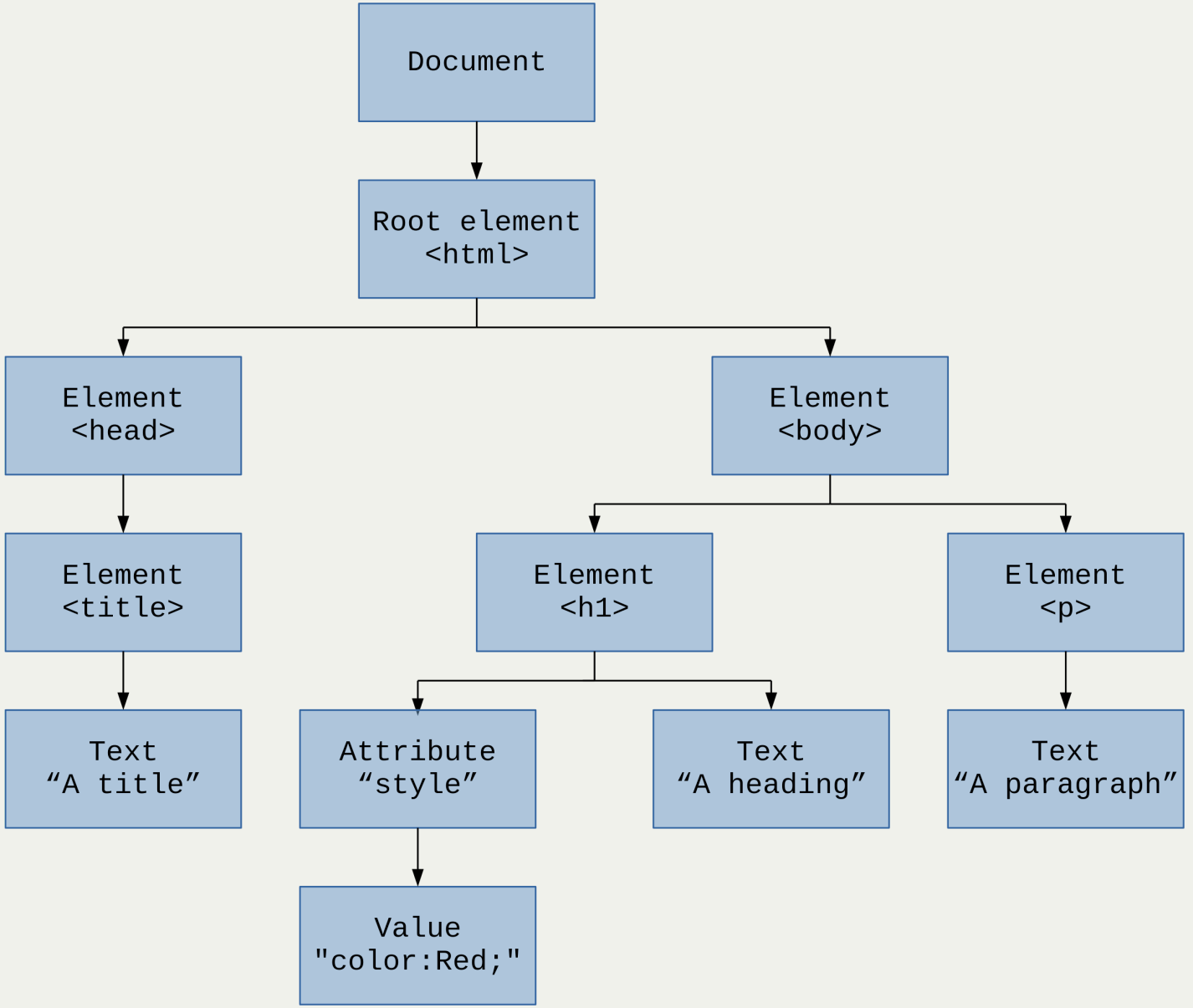
Extra

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 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;"`)
- Soma HTML tag exampes:
  - Document structure: `<html>`, `<body>`, `<header>`
  - Document components: `<h1>`, `<title>`, `<div>`
  - Text style: `<b>`, `<i>`
  - Hyperlinks: `<a>`

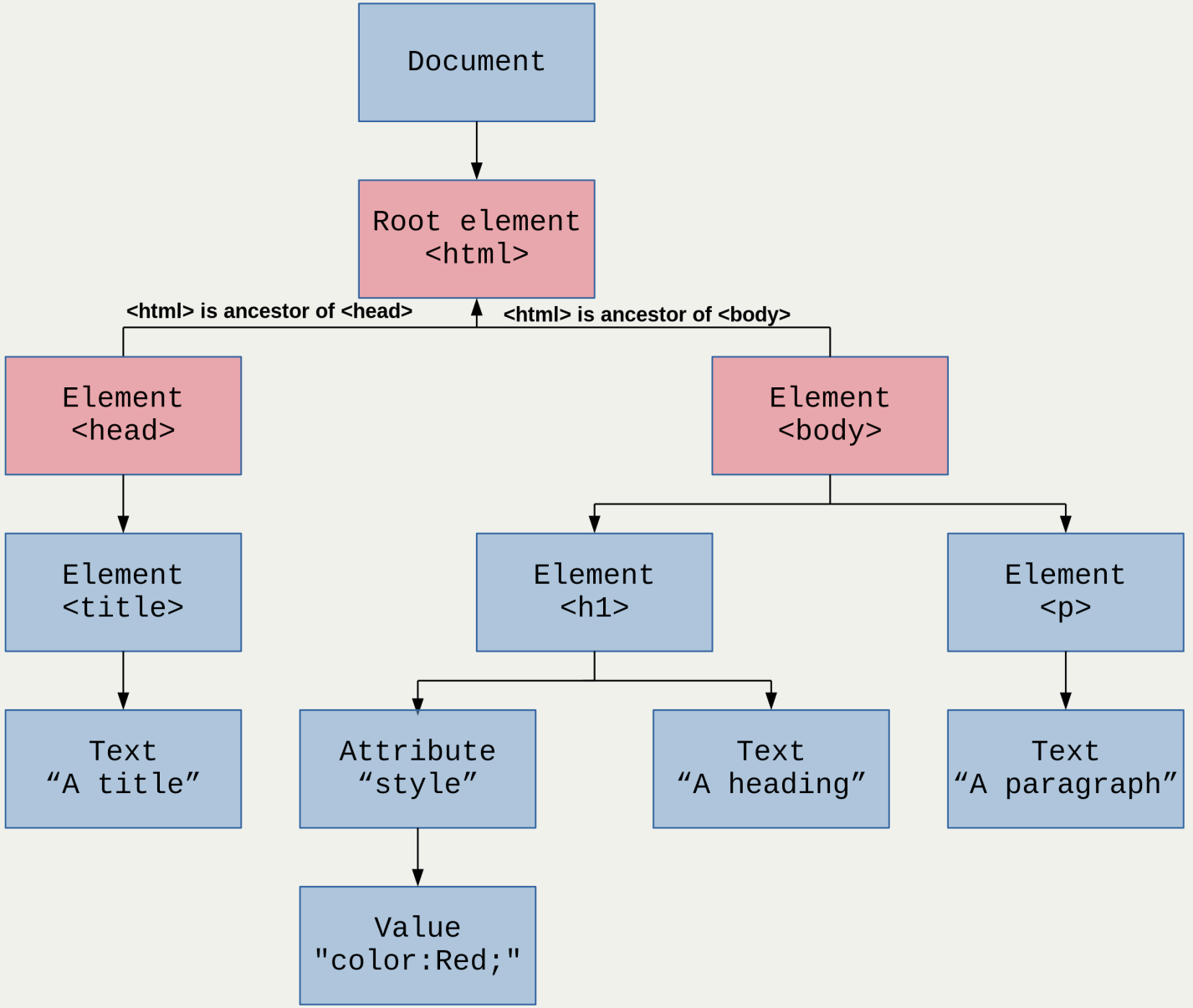
HTML tree


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;"

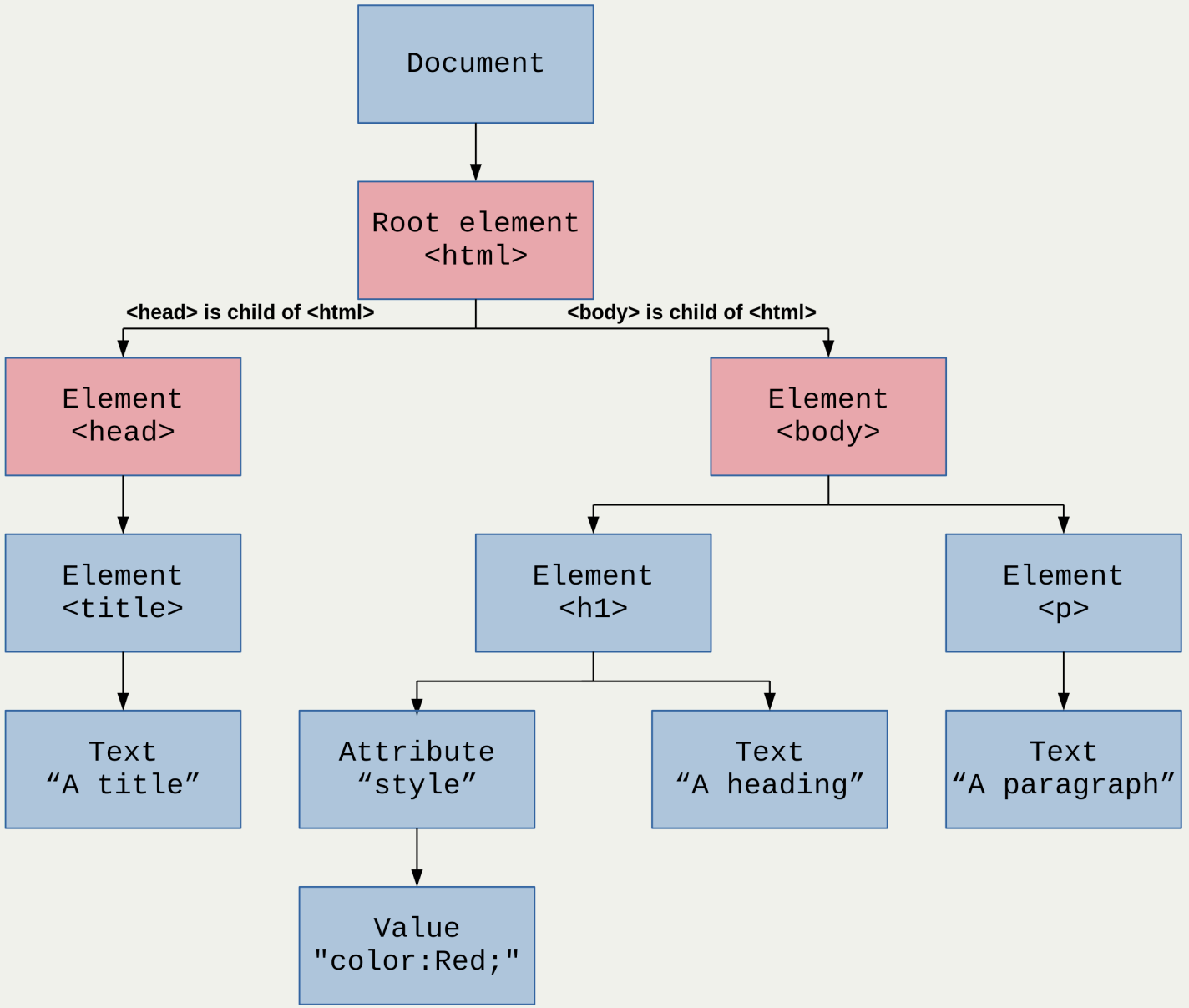
HTML Tree Relationships

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

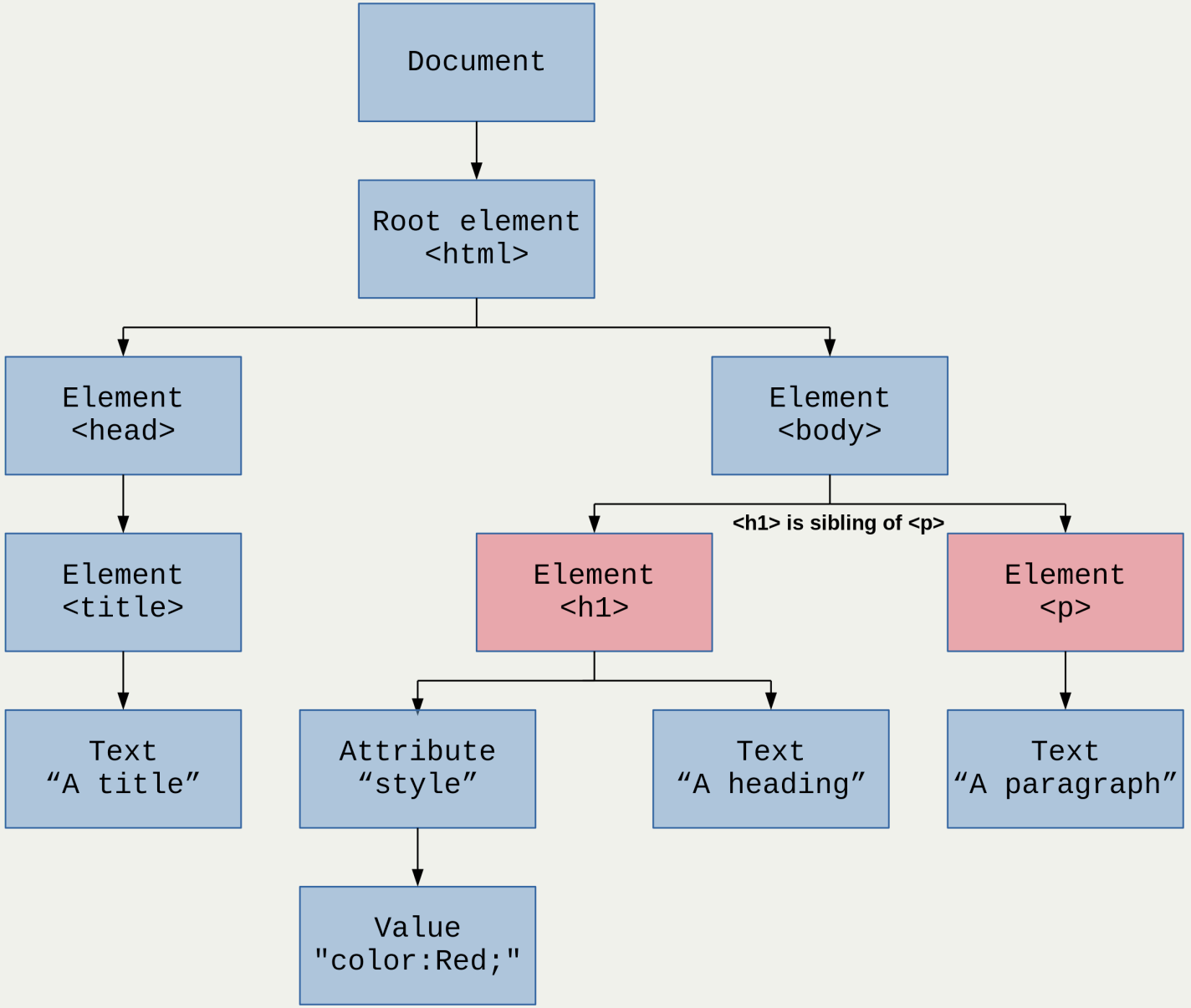
HTML Parent/Ancestor



HTML Children/Descendants



HTML Siblings



Parsing HTML Tree: Example

```r
library("rvest")
```

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

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

```r
str(html)
```

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

Parsing HTML Tree: Example

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

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

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

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

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

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

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

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

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

XML: Extensible Markup Language

- XML (**Ex**tensible **M**arkup **L**anguage) 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

```r
library("xml2")
```

```r
xml_txt <-
'<?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>'
```

```r
xml <- xml2::read_xml(xml_txt)
```

```r
str(xml)
```

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

Parsing XML Tree: Example

```r
children3 <- xml2::xml_children(xml)
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 ...
```

```r
pop77001 <- children3[1]
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>
```

```r
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

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

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

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

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

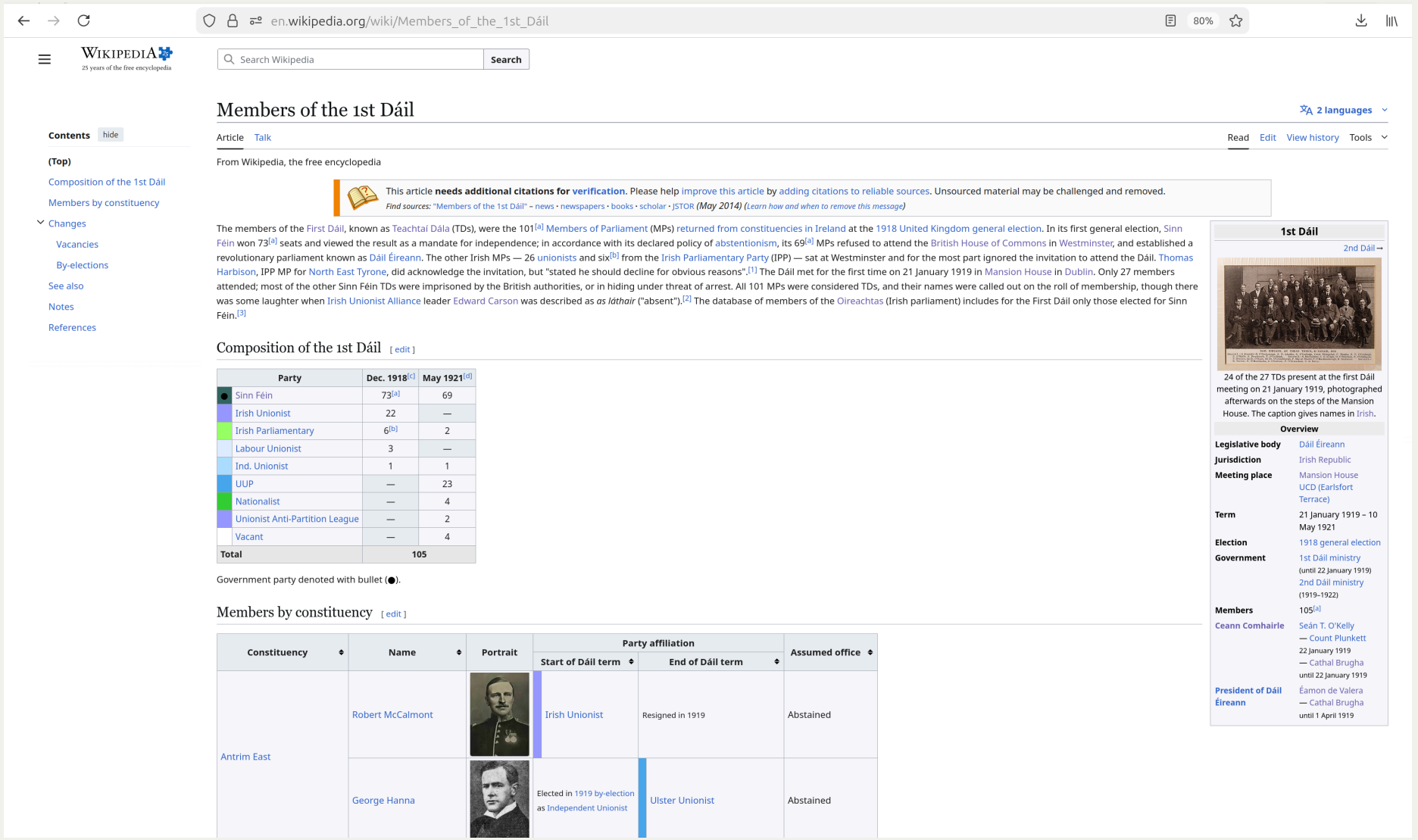
```
xml2::xml_find_all(xml, xpath = "//code")
```

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

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

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

Scraping Webpage



(Wikipedia)

Scraping Webpage with XPath: Example

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

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

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

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

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

```
 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-S… ""       ""                  Irish Unionist
 6 Antrim South   Charles Craig ""       ""                  Irish Unionist
 7 Armagh Mid     James Lonsda… ""       ""                  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 Craig    ""       ""                  Irish Unionist
```

Scraping Webpage with XPath: Example

```r
colnames(tds) <- tds[1,]
tds <- tds[-1,]
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 Craig  ""       ""                   Irish Unionist
6 Armagh Mid   James Lonsdale ""       ""                   Irish Unionist
# ℹ 3 more variables: `End of Dáil term` <chr>, `End of Dáil term` <chr>,
#   `Assumed office` <chr>
```

```r
tds <- tds[,-3]
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