Statistical Programming and Open Science Methods

Tidy data scraping

Joachim Gassen
Humboldt-Universität zu Berlin

September 02, 2022

SFB/Transregio 266
ACCOUNTING FOR TRANSPARENCY

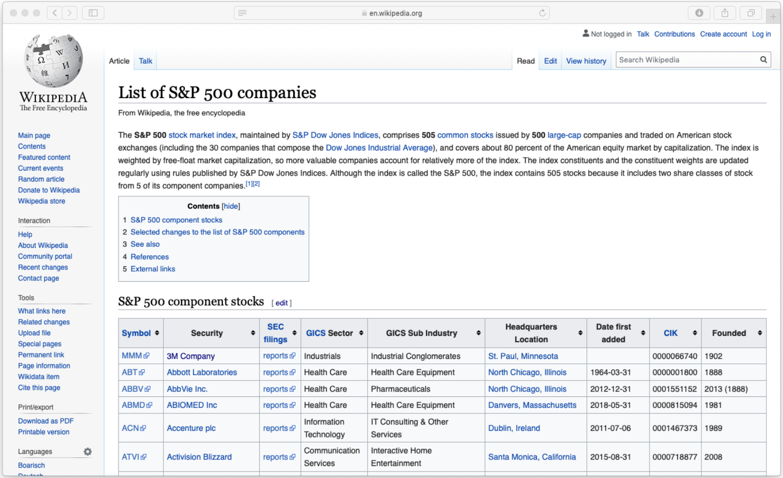
Time table Monday, February 17

When?	What?
10:00	Welcome and Coffee
10:30	Tidy data scraping
12:00	Lunch
13:30	Code along: Unit testing in function development
15:00	Coffee
15:30	Group work presentations
16:30	End of Day

Time table Tuesday, February 18

When?	What?
09:00	Explore your researcher degrees of freedom
10:30	Coffee
11:00	Providing data access via RESTful APIs
12:30	Lunch
13:30	Group Work Presentations
15:00	Coffee and Wrap Up
15:30	End of Event

Parsing an HTML table I



List of S&P 500 companies

From Wikipedia, the free encyclopedia

The **S&P 500** stock market index, maintained by S&P Dow Jones Indices, comprises **505** common stocks issued by **500** large-cap companies and traded on American stock exchanges (including the 30 companies that compose the Dow Jones Industrial Average), and covers about 80 percent of the American equity market by capitalization. The index is weighted by free-float market capitalization, so more valuable companies account for relatively more of the index. The index constituents and the constituent weights are updated regularly using rules published by S&P Dow Jones Indices. Although the index is called the S&P 500, the index contains 505 stocks because it includes two share classes of stock from 5 of its component companies.[1][2]

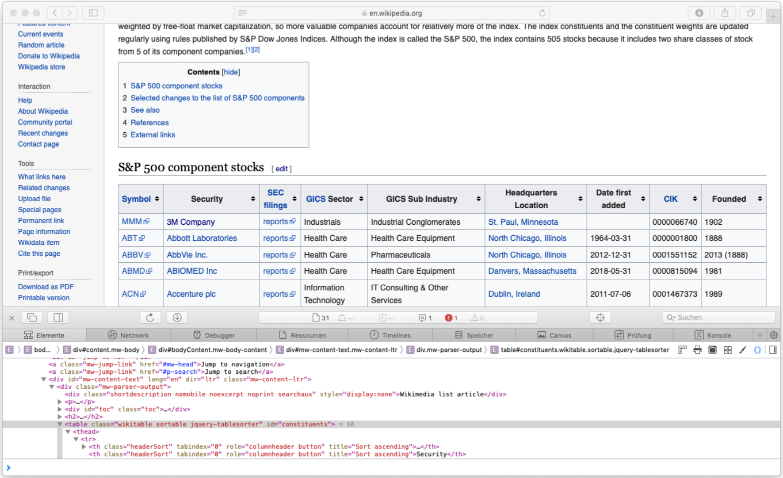
Contents [hide]

1 S&P 500 component stocks
2 Selected changes to the list of S&P 500 components
3 See also
4 References
5 External links

S&P 500 component stocks [edit]

Symbol	Security	SEC filings	GICS Sector	GICS Sub Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota		0000066740	1902
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV	AbbVie Inc.	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ABMD	ABIOMED Inc	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989
ATVI	Activision Blizzard	reports	Communication Services	Interactive Home Entertainment	Santa Monica, California	2015-08-31	0000718877	2008

Parsing an HTML table II



weighted by free-float market capitalization, so more valuable companies account for relatively more of the index. The index constituents and the constituent weights are updated regularly using rules published by S&P Dow Jones Indices. Although the index is called the S&P 500, the index contains 505 stocks because it includes two share classes of stock from 5 of its component companies.[1][2]

Contents [hide]

1 S&P 500 component stocks
2 Selected changes to the list of S&P 500 components
3 See also
4 References
5 External links

S&P 500 component stocks [edit]

Symbol	Security	SEC filings	GICS Sector	GICS Sub Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota		0000066740	1902
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV	AbbVie Inc.	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ABMD	ABIOMED Inc	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989

```
      <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
      <a class="mw-jump-link" href="#p-search">Jump to search</a>
    ▼ <div id="mw-content-text" lang="en" dir="ltr" class="mw-content-ltr">
      ▼ <div class="mw-parser-output">
          <div class="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">Wikimedia list article</div>
        ▶ <p>…</p>
        ▶ <div id="toc" class="toc">…</div>
        ▶ <h2>…</h2>
        ▼ <table class="wikitable sortable jquery-tablesorter" id="constituents"> = $0
          ▼ <thead>
            ▼ <tr>
              ▶ <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">…</th>
                <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Security</th>
```

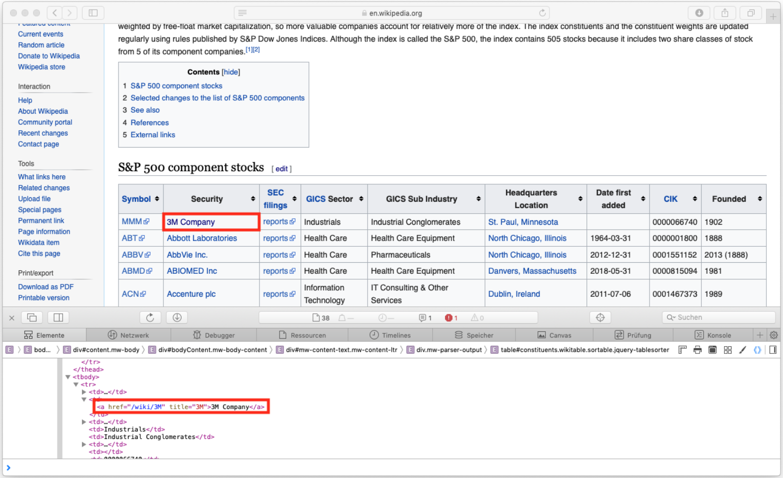
Parsing an HTML table III

```r
library(tidyverse)
library(rvest)

url_sp500_const <- paste0(
  "https://en.wikipedia.org/wiki/",
  "List_of_S%26P_500_companies"
)

url_sp500_const %>%
  read_html() %>%
  html_node(xpath = '//*[@id="constituents"]') %>%
  html_table() -> sp500_constituents_raw
```

Retrieving local URLs from within table I



weighted by free-float market capitalization, so more valuable companies account for relatively more of the index. The index constituents and the constituent weights are updated regularly using rules published by S&P Dow Jones Indices. Although the index is called the S&P 500, the index contains 505 stocks because it includes two share classes of stock from 5 of its component companies.[1][2]

Contents [hide]

1 S&P 500 component stocks
2 Selected changes to the list of S&P 500 components
3 See also
4 References
5 External links

S&P 500 component stocks [edit]

Symbol	Security	SEC filings	GICS Sector	GICS Sub Industry	Headquarters Location	Date first added	CIK	Founded
MMM	3M Company	reports	Industrials	Industrial Conglomerates	St. Paul, Minnesota		0000066740	1902
ABT	Abbott Laboratories	reports	Health Care	Health Care Equipment	North Chicago, Illinois	1964-03-31	0000001800	1888
ABBV	AbbVie Inc.	reports	Health Care	Pharmaceuticals	North Chicago, Illinois	2012-12-31	0001551152	2013 (1888)
ABMD	ABIOMED Inc	reports	Health Care	Health Care Equipment	Danvers, Massachusetts	2018-05-31	0000815094	1981
ACN	Accenture plc	reports	Information Technology	IT Consulting & Other Services	Dublin, Ireland	2011-07-06	0001467373	1989

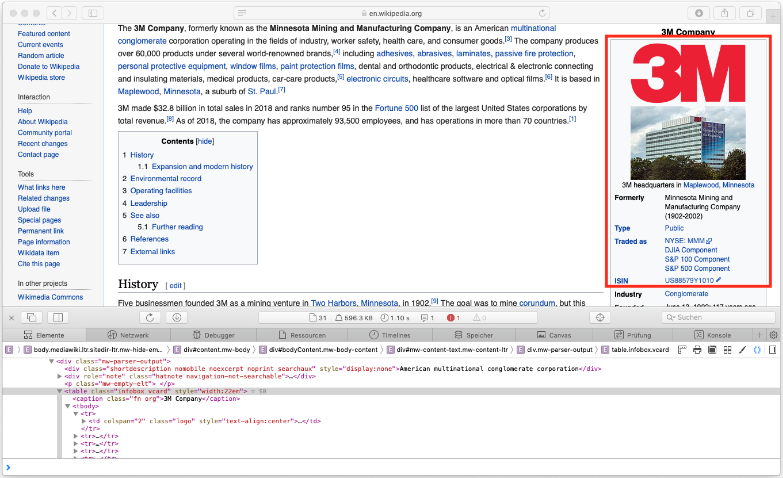
bod… 〉 div#content.mw-body 〉 div#bodyContent.mw-body-content 〉 div#mw-content-text.mw-content-ltr 〉 div.mw-parser-output 〉 table#constituents.wikitable.sortable.jquery-tablesorter

```
        </tr>
      </thead>
    ▼ <tbody>
      ▼ <tr>
        ▶ <td>…</td>
        ▼ <td>
            <a href="/wiki/3M" title="3M">3M Company</a>
          </td>
        ▶ <td>…</td>
          <td>Industrials</td>
          <td>Industrial Conglomerates</td>
        ▶ <td>…</td>
          <td></td>
```

Retrieving local URLs from within table II

```
url_sp500_const %>%
  read_html() %>%
  html_node(xpath = '//*[@id="constituents"]') %>%
  html_nodes("td:nth-child(2) a") %>%
  html_attr("href")-> links
```

Scraping ill-structured tables I



The **3M Company**, formerly known as the **Minnesota Mining and Manufacturing Company**, is an American multinational conglomerate corporation operating in the fields of industry, worker safety, health care, and consumer goods.[3] The company produces over 60,000 products under several world-renowned brands,[4] including adhesives, abrasives, laminates, passive fire protection, personal protective equipment, window films, paint protection films, dental and orthodontic products, electrical & electronic connecting and insulating materials, medical products, car-care products,[5] electronic circuits, healthcare software and optical films.[6] It is based in Maplewood, Minnesota, a suburb of St. Paul.[7]

3M made $32.8 billion in total sales in 2018 and ranks number 95 in the Fortune 500 list of the largest United States corporations by total revenue.[8] As of 2018, the company has approximately 93,500 employees, and has operations in more than 70 countries.[1]

Contents [hide]

1 History
1.1 Expansion and modern history
2 Environmental record
3 Operating facilities
4 Leadership
5 See also
5.1 Further reading
6 References
7 External links

History [edit]

Five businessmen founded 3M as a mining venture in Two Harbors, Minnesota, in 1902.[9] The goal was to mine corundum, but this

3M Company

3M headquarters in Maplewood, Minnesota

Formerly	Minnesota Mining and Manufacturing Company (1902-2002)
Type	Public
Traded as	NYSE: MMM DJIA Component S&P 100 Component S&P 500 Component
ISIN	US88579Y1010
Industry	Conglomerate

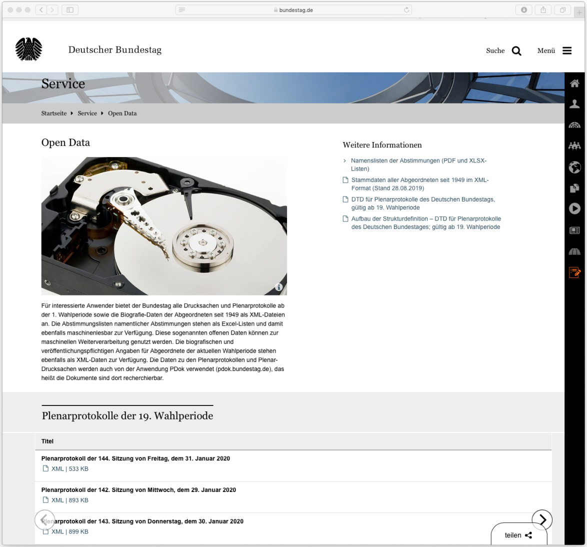
```
▼<div class="mw-parser-output">
   <div class="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">American multinational conglomerate corporation</div>
  ▶<div role="note" class="hatnote navigation-not-searchable">…</div>
   <p class="mw-empty-elt"> </p>
  ▼<table class="infobox vcard" style="width:22em"> = $0
     <caption class="fn org">3M Company</caption>
    ▼<tbody>
      ▼<tr>
        ▶<td colspan="2" class="logo" style="text-align:center">…</td>
       </tr>
      ▶<tr>…</tr>
      ▶<tr>…</tr>
      ▶<tr>…</tr>
```

Scraping ill-structured tables II

```r
xml_data <- read_html(url) %>%
  html_node('#mw-content-text div table.infobox.vcard')

xml_data  %>%
  html_table(fill = TRUE) %>%
  rename(tag = X1, content = X2) %>%
  filter(tag != "")
```

Another approach to scraping



Service

Open Data

Für interessierte Anwender bietet der Bundestag alle Drucksachen und Plenarprotokolle ab der 1. Wahlperiode sowie die Biografie-Daten der Abgeordneten seit 1949 als XML-Dateien an. Die Abstimmungslisten namentlicher Abstimmungen stehen als Excel-Listen und damit ebenfalls maschinenlesbar zur Verfügung. Diese sogenannten offenen Daten können zur maschinellen Weiterverarbeitung genutzt werden. Die biografischen und veröffentlichungspflichtigen Angaben für Abgeordnete der aktuellen Wahlperiode stehen ebenfalls als XML-Daten zur Verfügung. Die Daten zu den Plenarprotokollen und Plenar-Drucksachen werden auch von der Anwendung PDok verwendet (pdok.bundestag.de), das heißt die Dokumente sind dort recherchierbar.

Weitere Informationen

- Namenslisten der Abstimmungen (PDF und XLSX-Listen)
- Stammdaten aller Abgeordneten seit 1949 im XML-Format (Stand 28.08.2019)
- DTD für Plenarprotokolle des Deutschen Bundestags, gültig ab 19. Wahlperiode
- Aufbau der Strukturdefinition – DTD für Plenarprotokolle des Deutschen Bundestages; gültig ab 19. Wahlperiode

Plenarprotokolle der 19. Wahlperiode

Titel
Plenarprotokoll der 144. Sitzung von Freitag, dem 31. Januar 2020 XML | 533 KB
Plenarprotokoll der 142. Sitzung von Mittwoch, dem 29. Januar 2020 XML | 893 KB
Plenarprotokoll der 143. Sitzung von Donnerstag, dem 30. Januar 2020 XML | 899 KB

This won't work

```
library(rvest)

url_bt_open_data <- "https://www.bundestag.de/services/opendata"
url_bt_open_data %>%
  read_html() %>%
  html_node(
    xpath = '//*[@id="bt-collapse-543410"]/div[1]/div/div/div[1]/table'
  ) %>% html_table() -> pp_table
```

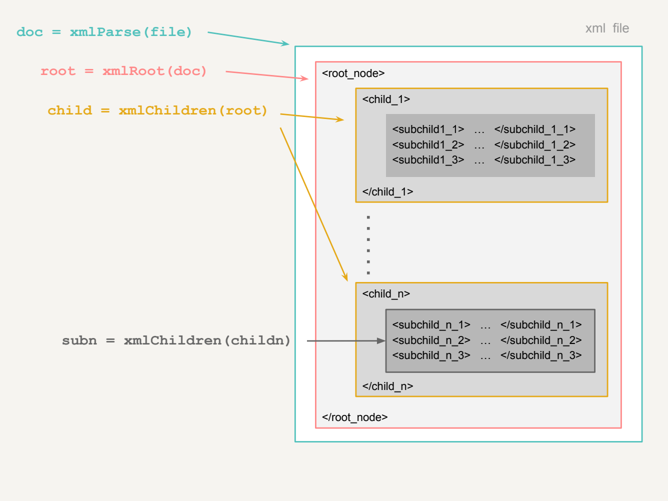
```
## Error in UseMethod("html_table"): no applicable method for 'html_tab
```

Reason: The web page is being created dynamically by JavaScript (or similar)

The idea of headless browsing: meet Selenium

- Selenium offers a way to script a web browser so that data can be scraped using the navigation that a web page provides
- Allows for various web browser and all sorts of user web browser interaction
- Uses a docker container holding the actual browser environment
- See `code/btag_open_data_scrape_data.R` for a demonstration on how to use it

Parsing XML Data



xml file
doc = xmlParse(file)
root = xmlRoot(doc)
child = xmlChildren(root)
subn = xmlChildren(childn)
<root_node>
<child_1>
<subchild1_1> … </subchild_1_1>
<subchild1_2> … </subchild_1_2>
<subchild1_3> … </subchild_1_3>
</child_1>
<child_n>
<subchild_n_1> … </subchild_n_1>
<subchild_n_2> … </subchild_n_2>
<subchild_n_3> … </subchild_n_3>
</child_n>
</root_node>

Gaston Sanchez, https://github.com/gastonstat/tutorial-R-web-data

My task . . .

- Develop a function that parses the XML files of the Plenarprotokolle of the 19th Wahlperiode to extract all speaches into a tidy data structure
- Implement some basic test routines verifying that the code works
- See `code/btag_open_data_scrape_data.R` and `code/test/test_btag_open_data_scrape_data.R`