

Web scraping for drug safety

R-thritis Computing Group

David A. Selby

5th November 2021

Structure

1. Why Web scraping?
2. Intro to HTML/CSS
3. Web scraping with rvest

Why Web scraping?

Why Web scraping?

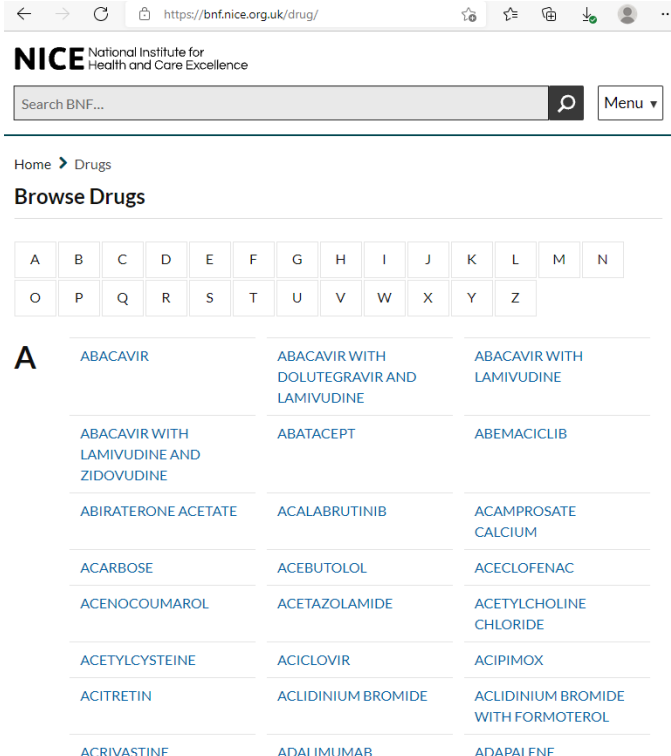
- > There's lots of useful information online
- > Not everything is a CSV file!
- > Faster / less error-prone than copying data manually
- > Fun

Motivating example

BNF

British National Formulary

- > <https://bnf.nice.org.uk/drug/>
- > One page per drug
- > Drug dose indications



The screenshot shows the NICE website interface. At the top, there is a search bar labeled 'Search BNF...' and a 'Menu' dropdown. Below the search bar, the navigation path 'Home > Drugs' is visible. The main heading is 'Browse Drugs'. A grid of letters from A to Z is displayed, with 'A' selected. Below the grid, a list of drug names is shown in a three-column format under the heading 'A'. The drugs listed include ABACAVIR, ABACAVIR WITH DOLUTEGRAVIR AND LAMIVUDINE, ABACAVIR WITH LAMIVUDINE, ABEMACICLIB, ABACAVIR WITH LAMIVUDINE AND ZIDOVUDINE, ABATACEPT, ABIRATERONE ACETATE, ACALABRUTINIB, ACAMPROSATE CALCIUM, ACARBOSE, ACEBUTOLOL, ACECLOFENAC, ACENOCOUMAROL, ACETAZOLAMIDE, ACETYLCHOLINE CHLORIDE, ACETYLCYSTEINE, ACICLOVIR, ACIPIMOX, ACITRETIN, ACLIDINIUM BROMIDE, ACLIDINIUM BROMIDE WITH FORMOTEROL, ACRIVASTINE, ADALIMUMAB, and ADAPALENE.

HTML for dummies

Example HTML document

```
<HTML>
  <HEAD>
    <TITLE>The title of my Web page</TITLE>
  </HEAD>
  <BODY>
    <H1>A heading</H1>
    <P>A paragraph about something.</P>
    <P>A second paragraph about something <em>else</em></P>
    <IMG SRC="logo.jpg" ALT="CfE logo">
    <UL> <!-- This is an unordered list -->
      <LI>A <A HREF="https://cfe.manchester.ac.uk">hyperlink</A>.
      <LI>Another list item</LI>
    </UL>
  </BODY>
</HTML>
```

Example HTML document

Example HTML document

```
<HTML>
  <HEAD>
    <TITLE>The title of my Web page</TITLE>
  </HEAD>
  <BODY>
    <H1 ID="headline">A heading</H1>
    <P CLASS="intro">A paragraph about something.</P>
    <P>A second paragraph about something <em>else</em></P>
    <IMG SRC="logo.jpg" ALT="CfE logo" CLASS="logo">
    <UL> <!-- This is an unordered list -->
      <LI>A <A HREF="https://cfe.manchester.ac.uk">hyperlink</A>.
      <LI>Another list item</LI>
    </UL>
  </BODY>
</HTML>
```

Cascading style sheets (CSS)

Use **tags**, **classes** and **ids** to identify objects in the DOM.

e.g. Select the headline text:

- > `h1`
- > `h1#headline` (or `#headline`)
- > `body:first-child`

e.g. Select the introduction paragraph:

- > `p.intro` (or `.intro`)
- > `p:first-of-type`
- > `h1+p`
- > `body:nth-child(2)`

Cascading style sheets (CSS)

Style:

- [1] change the typeface
- [2] centre the headline
- [3] highlight the intro paragraph
- [4] shrink the logo image

Add the following in `<style>` `</style>` tags:

```
body { font-family: 'Comic Sans MS'; }  
h1#headline { text-align: center; }  
.intro { background-color: yellow; }  
.logo { width: 100px; }
```

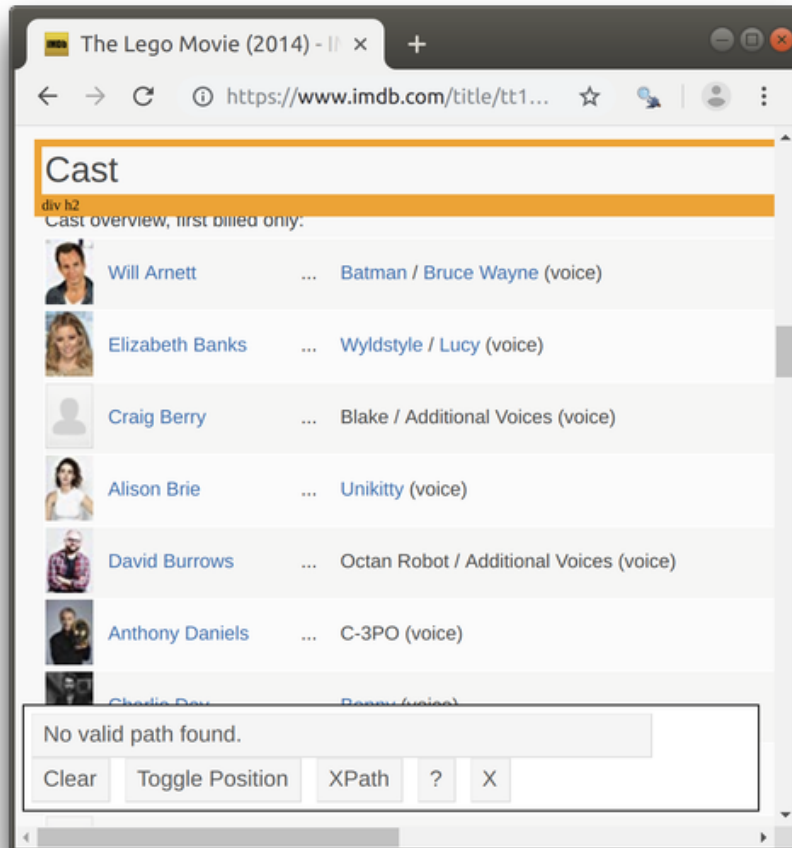
Example HTML document with CSS

The element inspector

Explore the document object model (DOM) of any Web page:

SelectorGadget

<https://rvest.tidyverse.org/articles/selectorgadget.html>



Web scraping with rvest

Web scraping with rvest

```
library(rvest)
example <- read_html('example.html')
```

```
# {html_document}
# <html>
# [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">\n</head>\n<body>\r\n  <h1 id="headline">A heading</h1>\r\n  <p class="intro">A paragraph about something.\r\n</body>\n</html>
```

```
example %>% html_element('.intro')
```

```
# {html_node}
# <p class="intro">
```

```
example %>% html_element('.intro') %>% html_text()
```

```
# [1] "A paragraph about something."
```


Web scraping with rvest

```
drug_index <- read_html('https://bnf.nice.org.uk/drug/')
drug_links <- drug_index %>% html_elements('.row ul li a')
drugs <- data.frame(name = html_text2(drug_links),
                    path = html_attr(drug_links, 'href'))
head(drugs)
```

```
# A tibble: 6 x 2
```

name	path
<chr>	<chr>
1 ABACAVIR	abacavir.html
2 ABACAVIR WITH DOLUTEGRAVIR AND LAMIVUDINE	abacavir-with-dolutegravir-and-
3 ABACAVIR WITH LAMIVUDINE	abacavir-with-lamivudine.html
4 ABACAVIR WITH LAMIVUDINE AND ZIDOVUDINE	abacavir-with-lamivudine-and-zi
5 ABATACEPT	abatacept.html
6 ABEMACICLIB	abemaciclib.html

Web scraping with rvest

```
library(tidyverse)
scrape_drug <- function(path) {
  webpage <- read_html(file.path('https://bnf.nice.org.uk/drug/',
  name_of_drug <- webpage %>% html_element('h1') %>% html_text2
  condition_grp <- webpage %>% html_elements('.indicationAndDoseGr
  condition_name <- map(condition_grp, ~ html_element(.x, '.indica
  tibble(name_of_drug,
          condition = map_chr(condition_name, paste, collapse = ', ')
          route_grp = map(condition_grp, html_elements, '.c
  unnest(route_grp) %>%
  mutate(route = map_chr(route_grp, ~ html_elements(.x, 'span.rc
          patient_grp = map(route_grp, html_elements, 'li.dose')
  unnest(patient_grp) %>%
  mutate(patient = map_chr(patient_grp, ~ html_element(.x, '.pat
          dose = map_chr(patient_grp, ~ html_elements(.x, 'p') %>%
  select(-ends_with('_grp'))
}
```

Ibuprofen example

```
scrape_drug('ibuprofen.html')
```

```
# # A tibble: 24 x 5
#   name_of_drug condition                route      patient      dose
#   <chr>          <chr>                <chr>      <chr>      <chr>
# 1 IBUPROFEN     "Pain and inflammatio~ By mouth us~ Adult      Initially 3
# 2 IBUPROFEN     "Pain and inflammatio~ By mouth us~ Adult      1.6&nbsp;g
# 3 IBUPROFEN     "Acute migraine\n"    By mouth us~ Adult      400-600&nbsp;
# 4 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 3-~ 50&nbsp;mg
# 5 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 6-~ 50&nbsp;mg
# 6 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 1-~ 100&nbsp;mg
# 7 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 4-~ 150&nbsp;mg
# 8 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 7-~ 200&nbsp;mg
# 9 IBUPROFEN     "Mild to moderate pai~ By mouth us~ Child 10~ 300&nbsp;mg
# 10 IBUPROFEN    "Mild to moderate pai~ By mouth us~ Child 12~ Initially 3
# # ... with 14 more rows
```

More information

- > <https://rvest.tidyverse.org>
- > Blog post: *'Which film should I watch during lockdown?'*
<https://selbydavid.com>
- > E-mail me: david.selby@manchester.ac.uk

Upcoming R-thritis meetings

19 November ←

Topic/presenter to be confirmed

3 December ←

'Advent of Code' discussion