

HARVESTING ONLINE DATA

Short course at Santa Clara University, ECON173

Johannes Fritz

5/5/22, 5/12/22

WHY AM I HERE?

- ▶ Economist with a strong “data science tilt”
 - > Web harvesting
 - > Machine learning
 - > Natural language processing
- ▶ Data collection is my day job
 - > Global Trade Alert
Dataset on industrial policy choices around the world since 2009
 - > Digital Policy Alert
Dataset on G20+Europe’s digital policy choices since 2021

YOU WILL LEARN 3 TOOLS

1. Harvesting the data:
Headless **web scraping** using R.

today

2. Parsing the data:
Introduction to **Regular Expressions**.

next

3. Organizing the data:
Basics of **relational database design**.

week

All material at:

<https://github.com/johannesfritz/scu-harvesting>

GETTING STARTED

- ▶ What projects do you want to tackle?
- ▶ What experience do you already have?
- ▶ What tools do you want to learn?

THE U.S. CONGRESSIONAL RECORD

MY TEACHING EXAMPLE FOR THIS SHORT COURSE

THE U.S. CONGRESSIONAL HEARINGS

Source

- ▶ <https://www.govinfo.gov/app/collection/CHRG/>
- ▶ or search:
«GPO Congressional Hearings»

govinfo

[Browse](#)[About](#)[Developers](#)[Features](#)[Help](#)[Feedback](#)

Q SEARCH ▾

Congressional Hearings

[f](#)[t](#)[e](#)[s](#)[r](#)

HELP

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. [Read More](#)

Browse Congressional Hearings

117th Congress (2021 - 2022)

House Hearings

Committee on Agriculture

Committee on Appropriations

Committee on Armed Services

Committee on Education and Labor

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

General. Thursday, March 11, 2021.

PDF

TEXT

DETAILS

SHARE

Sort **↓**_A^Z

WEBSITE CONTENTS

govinfo

[Browse](#)[About](#)[Developers](#)[Features](#)[Help](#)[Feedback](#)

Q SEARCH ▴

Download

PDF

Text

MODS

PREMIS

ZIP

Actions

Browse Congressional Hearings

Help

CGP Record

Content Details

House Hearing, 117th Congress - CLEARING THE AIR: SCIENCE-BASED STRATEGIES TO PROTECT WORKERS FROM COVID-19 INFECTIONS

Summary

Document in Context ⓘ

Category

Congressional Committee Materials

Collection

Congressional Hearings

SuDoc Class Number

Y 4.ED 8/1:117-1

ILS System ID

001171899

Congress

117th

Congress Chamber

House of Representatives

Hearing Sub Type

General

Held Date

March 11, 2021

Committee and Subcommittee

Committee on Education and Labor

Members

Joe Wilson (SC); Raul M. Grijalva (AZ); Virginia Foxx (NC); Joe Courtney (CT); John A. Yarmuth (KY); Tim Walberg (MI); Glenn Thompson (PA); Frederica S. Wilson (FL); Suzanne Bonamici (OR); Mark Takano (CA); Joaquin Castro (TX); Mark Pocan (WI); Alma S. Adams (NC); Donald Norcross (NJ); Rick W. Allen (GA); Elise M. Stefanik (NY); Glenn Grothman (WI); James Comer (KY); Jim Banks (IN); Adriano Espaillat (NY); Pramila Jayapal (WA); Joseph D. Morelle (NY); Susan Wild (PA); Jahana Hayes (CT); Russ Fulcher (ID); Andy Levin (MI); Haley M. Stevens (MI); Ilhan Omar (MN); Mikie Sherrill (NJ); Fred Keller (PA); Gregory Murphy (NC); Michelle Steel (CA); Mariannette Miller-Meeks (IA); Mary E. Miller (IL); Frank J. Mrvan (IN); Victoria Spartz (IN); Kathy E. Manning (NC); Madison Cawthorn (NC); Teresa Leger Fernandez (NM); Jamaal Bowman (NY); Mondaire Jones (NY); Diana Harshbarger (TN); Burgess Owens (UT); Bob Good (VA); Scott Fitzgerald (WI); Kweisi Mfume (MD)

Witnesses

Marr, Linsey Ph.D., Professor of Civil and Environmental Engineering, Virginia Polytechnical Institute and State University, Blacksburg, VA; Michaels, David, Ph.D., Professor of Occupational and Environmental Medicine, The George Washington University, Former Assistant Secretary of OSHA, Washington, DC; Muhindura, Pascaline, RN, COVID Progressive Care Unit, Research Medical Center, on behalf of National Nurses United, Kansas City, MO; Rath, Manesh, Partner, Keller and Heckman LLP, Washington, DC

Serial Numbers

Serial No. 117-1

Bill Numbers

H.R. 1180
H.R. 1195
H.R. 1319

[House Hearing, 117 Congress]
[From the U.S. Government Publishing Office]

CLEARING THE AIR: SCIENCE-BASED STRATEGIES
TO PROTECT WORKERS FROM
COVID 19 INFECTIONS

=====

HEARING

BEFORE THE

SUBCOMMITTEE ON
WORKFORCE PROTECTIONS

OF THE

COMMITTEE ON EDUCATION AND LABOR
U.S. HOUSE OF REPRESENTATIVES

ONE HUNDRED SEVENTEENTH CONGRESS

FIRST SESSION

HEARING HELD IN WASHINGTON, DC, MARCH 11, 2021

Serial No. 117-1

BUT:

HEARINGS ARE NOT RESEARCH READY

No search, nor filtered export

- > Dates, topics
- > Members of Congress, Witnesses
- > Bills
- > Individual speeches

Have to create the dataset ourselves!

1. Harvest it.
2. Parse it.
3. Store it.

YOU WILL LEARN 3 TOOLS

1. Harvesting the data:
Headless **web scraping** using R.
2. Parsing the data:
Introduction to **Regular Expressions**.
3. Organizing the data:
Basics of **relational database design**.

All material at:

<https://github.com/johannesfritz/scu-harvesting>

I.

HARVESTING THE DATA

HEADLESS WEB SCRAPING USING R

TWO APPROACHES TO HARVESTING WEBSITES

1. GET requests

- > Effectively downloading the HTML/etc. files
- > R library {httr}

2. Browser automation + GET requests

- > Used for website testing
- > «headless»: without open window (GUI)
- > R library {webdriver}

WHY WE NEED BROWSER AUTOMATION.

- ▶ Don't know all the URLs.
- ▶ Site is dynamically rendered.

govinfo

[Browse](#)[About](#)[Developers](#)[Features](#)[Help](#)[Feedback](#)

Q SEARCH ▾

HELP

Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. [Read More](#)

Browse Congressional Hearings

117th Congress (2021 - 2022)


House Hearings

Committee on Agriculture

Committee on Appropriations

Committee on Armed Services

Committee on Education and Labor

Sort 

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

General. Thursday, March 11, 2021.

PDF

TEXT

DETAILS

SHARE

{WEBDRIVER} BASICS

- ▶ Library supporting «headless» browsing
 - > Using phantomJS
 - > Intuitive for rapid use
 - > Flexible extensions (if you know JavaScript)
 - > Alternative: {Rselenium}
- ▶ Supports
 - > Browsing
 - > Manipulating the page

{WEBDRIVER} BASICS

BROWSING

session = browser window

```
s <- Session$new(host = "127.0.0.1", port = 8910)
```

Log onto URL

```
s$delete()  
s$status()
```

like your browser icons

```
s$go(url)  
s$getUrl()  
s$goBack()  
s$goForward()  
s$refresh()  
s$getTitle()
```

extract HTML or screenshot

```
s$getSource()  
s$takeScreenshot(file = NULL)
```

{WEBDRIVER} BASICS

LOCATING + MANIPULATING TARGETS

search the HTML source code

```
e$findElement(css = NULL, linkText = NULL,  
  partialLinkText = NULL, xpath = NULL)  
e$findElements(css = NULL, linkText = NULL,  
  partialLinkText = NULL, xpath = NULL)
```

extract text (not code)

```
e$isSelected()  
e$getValue()  
e$setValue(value)  
e$getAttribute(name)  
e$getClass()  
e$getCssValue(name)  
e$getText()  
e$getName()  
e$getData(name)  
e$getRect()  
e$isEnabled()  
e$click()  
e$clear()  
e$sendKeys(...)
```

manipulate entry fields

OUR INITIAL SCRIPT

```
1 library(webdriver)
2 library(XML)
3
4 ## starting the headless browser
5 my.browser = run_phantomjs()
6 my.browser
7
8
9 # Opening a session and accessing the URL
10 my.url="https://www.govinfo.gov/app/collection/chrg/"
11 my.session = Session$new(port= my.browser$port)
12
13 my.session$go(my.url)
14
15
16 # Looking around the page
17
18 ## verifying I arrived at the URL
19 my.session$getUrl()
20 my.session$getTitle()
21
22 ## Taking screenshots
23 ## Very useful to check you're navigating right
24 my.session$takeScreenshot()
```

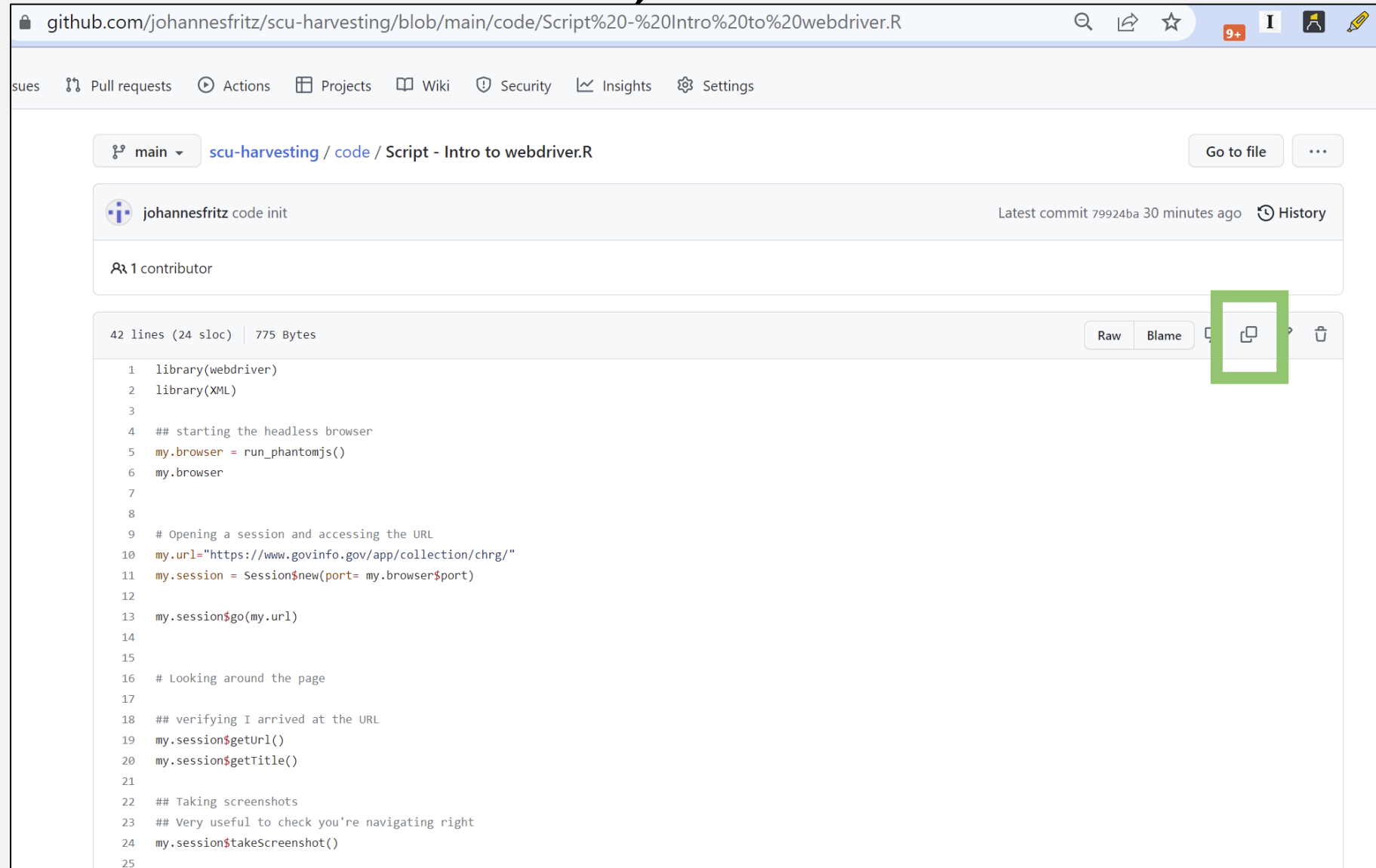

SUPER USEFUL: TAKING SCREENSHOTS



LOAD THE SCRIPT + SEE IF IT WORKS.

TRY OUT YOUR OWN PAGE, TOO

Copy it
from
GitHub



The screenshot shows a GitHub repository page for the file `Script - Intro to webdriver.R` in the `scu-harvesting` repository. The page includes a breadcrumb trail, repository navigation links, and a code editor view. A green box highlights the copy icon in the top right of the code editor.

```
1 library(webdriver)
2 library(XML)
3
4 ## starting the headless browser
5 my.browser = run_phantomjs()
6 my.browser
7
8
9 # Opening a session and accessing the URL
10 my.url="https://www.govinfo.gov/app/collection/chrg/"
11 my.session = Session$new(port= my.browser$port)
12
13 my.session$go(my.url)
14
15
16 # Looking around the page
17
18 ## verifying I arrived at the URL
19 my.session$getUrl()
20 my.session$getTitle()
21
22 ## Taking screenshots
23 ## Very useful to check you're navigating right
24 my.session$takeScreenshot()
25
```

MOVING ON:

KEY ELEMENTS OF OUR FULL SCRIPT

1. Navigate the site:
need to tell our browser where to click
> Tool: XPath
2. Create our local copy:
store all pages of interest
3. Make capture comprehensive:
go through all the pages we need
> Tool: for loops

[illegible]

HTML PRIMER

TAGS + ATTRIBUTES

Tag name	Function
<code><head></code>	The head of a page includes general information such as its authors, language etc.
<code><body></code>	All the content you see in the browser is inside the body tags.
<code><p></code>	A paragraph of text.
<code>
</code>	A break between two lines of text.
<code></code> , <code><i></code> , <code><u></code>	Display text in bold, italics or underlined.
<code></code>	A link to the page gpo.html.
<code></code>	Displays the picture gpo.jpg.
<code><table></code> , <code><tr></code> , <code><td></code>	A table including its rows (tr) and cells (td).
<code><div></code>	Divides an HTML document into different parts. Usually used for layout purposes.
<code></code>	Similar to <code><div></code> tag, but only for small parts of the code.

attribute

HTML PRIMER

DOCUMENT STRUCTURE

S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014

Appropriation. Wednesday, April 24, 2013.

[PDF](#) | [Text](#) | [More](#)

```
<table class="browse-node-table">
<tr id="this-is-the-1st-row">
<td colspan="2" id="this-is-the-only-cell"> S. Hrg. 113 - DEPARTMENT OF DEFENSE
APPROPRIATIONS FOR FISCAL YEAR 2014 </td> </tr>
<tr id="this-is-the-2nd-row">
<td id="this-is-one-cell"> <span class="results-line2"> Appropriation. Wednesday, April 24,
2013. </span> </td>
  <td id="this-is-another-cell"> <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
PDF </a> <a href="https://www.gpo.gov/link.to.text" target="_blank"> Text </a> <a
href="https://www.gpo.gov/link.to.more" target="_blank"> More </a> </td> </tr> <tr
id="this-is-the-3rd-row"> <td colspan="2" id="this-is-the-only-cell"> </td> </tr>
</table>
```

HTML PRIMER

DOCUMENT STRUCTURE - INDENTED

<div><table class="browse-node-table"></div> <div><tr id="this-is-the-1st-row"></div> <div><div><td colspan="2" id="this-is-the-only-cell"></div><div>S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014</div><div></td></div></div> <div></tr></div>	
<div><tr id="this-is-the-2nd-row"></div> <div><div><td id="this-is-one-cell"></div><div><div></div><div>Appropriation. Wednesday, April 24, 2013.</div><div></div></div><div></td></div></div> <td data-kind="ghost"></td>	
<div><td id="this-is-another-cell"></div> <div><div></div><div>PDF</div><div></div></div> <div><div></div><div>Text</div><div></div></div>	

indentation levels

- 1 →
- 2 →
- 3 →
- 4 →

XPATH

NAVIGATING THE DOM

- ▶ Simple query language to describe DOM locations.
- ▶ Basic rules:
 1. We only include the opening tags. The program will just ignore closing tags it sees inside the code along the way.
 2. The path only includes tags where the program has to move into a new nest (or towards the right in the above visualization).
 3. We use numbers in squared brackets to identify a specific tag in cases where there are multiple ones of the same kind within the same nest.
 4. We connect the different pieces of our path with backslashes in-between each tag.

XPATH EXAMPLE

CLICKING ON THE «TEXT» LINK

```
<table class="browse-node-table">
```

```
<tr id="this-is-the-1st-row">
```

```
<td colspan="2" id="this-is-the-only-cell">
```

```
S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
```

```
</td>
```

```
</tr>
```

```
<tr id="this-is-the-2nd-row">
```

```
<td id="this-is-one-cell">
```

```
<span class="results-line2">
```

```
Appropriation. Wednesday, April 24, 2013.
```

```
</span>
```

```
</td>
```

```
<td id="this-is-another-cell">
```

```
<a href="https://www.gpo.gov/link.to.PDF" target="_blank">
```

```
PDF
```

```
</a>
```

```
<a href="https://www.gpo.gov/link.to.text" target="_blank">
```

```
Text
```

```table```

```tr[2]```

```td[2]```

"Start at the beginning of our code (```table```),  
use the second row tag (```tr[2]```).

Within that row tag, go into the second cell tag you can find (```td[2]```), and  
```a[2]``` stop in that cell's second link tag (```a[2]```)."

➔ ```/table/tr[2]/td[2]/a[2]```

```
<table class="browse-node-table">
```

```
<tr id="this-is-the-1st-row">
```

```
<td colspan="2" id="this-is-the-only-cell">
```

S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014

```
</td>
```

```
</tr>
```

```
<tr id="this-is-the-2nd-row">
```

```
<td id="this-is-one-cell">
```

```
<span class="results-line2">
```

Appropriation. Wednesday, April 24, 2013.

```
</span>
```

```
</td>
```

```
<td id="this-is-another-cell">
```

```
<a href="https://www.gpo.gov/link.to.PDF" target="_blank">
```

PDF

```
</a>
```

```
<a href="https://www.gpo.gov/link.to.text" target="_blank">
```

Text

FINDING THE XPATH USING THE «INSPECTOR» IN CHROME/FIREFOX

Open with:

► CTRL + SHIFT + I

► ⌘ + SHIFT + I

► ⌘ + SHIFT + I

The screenshot shows the govinfo website with the Chrome DevTools 'Elements' panel open. The page title is 'Congressional Hearings'. The 'Elements' panel shows the HTML structure, with the 'body' element selected. The 'Styles' panel shows the CSS styles for the selected element, including 'padding-top: 65px;' and 'background-color: #fff;'. The 'Console' panel shows the 'What's New' message for Chrome 100.

govinfo

Browse About Developers Features Help Feedback

Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. [Read More](#)

Browse Congressional Hearings

117th Congress (2021 - 2022)

House Hearings

- Committee on Agriculture
- Committee on Appropriations
- Committee on Armed Services
- Committee on Education and Labor

Serial No. 117-1 (House Hearing) - Clearing the Path for Science-Based Strategies to Protect Workers from Covid-19 Infections

PDF TEXT DETAILS

SHARE

DevTools is now available in German! Always match Chrome's language Switch DevTools to German Don't show again

Elements Console Sources Network Performance Memory

```
<!DOCTYPE html>
<html lang="en">
  <head>...</head>
  <body data-new-gr-c-s-check-loaded="14.1058.0" data-gr-ext-installed="14.1058.0">
    <a href="#maincontent" class="skip-main">Skip to main content</a>
    <div id="collection" data-viewconstructor="CollectionView" class="fw-tab-toggle hide active" data-cid="view1"> </div>
    <div id="details" data-viewconstructor="NewContentDetailsTabView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="browse" data-viewconstructor="NewBrowseTabView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="search" data-viewconstructor="SearchView" class="fw-tab-toggle hide fw-tab-default hide" data-cid="view1"> </div>
    <div id="citation" data-viewconstructor="CitationView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="advanced" data-viewconstructor="AdvancedSearchView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="frtoc" data-viewconstructor="FrtocView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="cfrparts" data-viewconstructor="CFRPartsView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div id="today" data-viewconstructor="CCALView" class="fw-tab-toggle hide" data-cid="view1"> </div>
    <div role="main" id="homeId">
      <div id="homeheaderId" data-role="header">...</div>
      <div name="maincontent" id="maincontent" tabindex="-1"> </div>
      <div class="contentarea fw-tab-content" data-id="frtoc" data-role="content" data-cid="view1" style="display: none;"> </div>
      <div class="contentarea fw-tab-content" data-id="search" data-role="content" data-cid="view1" style="display: none;"> </div>
      <div class="contentarea fw-tab-content" data-id="browse" data-role="content" data-cid="view1" style="display: none;"> </div>
      <div class="contentarea fw-tab-content" data-id="collection" data-role="content" data-cid="view1" style="display: block;">
        <div id="browsecontentId">
          <div role="alert" class="col-xs-12 alert alert-warning hide" id="browseErrors"> </div>
          <div id="resultsId" aria-label="View The Results of your Browse here">
            <div tabindex="0" id="browsewidthId">

```

element.style {

```
@media screen and (min-width: 768px)
body {
  padding-top: 65px;
}
body {
  padding-top: 46px;
  background-color: #fff;
  height: 100%;
  color: #01203b;
  font-family: Helvetica, Arial, sans-serif;
  font-family: Roboto, sans-serif;
  font-size: 16px;
  line-height: 1.6;
  -webkit-font-smoothing: antialiased;
  -moz-osx-font-smoothing: grayscale;
}
body {
  font-family: Helvetica, Arial, sans-serif;
  font-size: 14px;
  line-height: 1.42857143;
  color: #333;
  background-color: #fff;
}
body {
  margin: 0;
}
* {
  -webkit-box-sizing: border-box;
  -moz-box-sizing: border-box;
  box-sizing: border-box;
}
body {
  display: block;
  margin: 0px;
}
Inherited from html
:root {
  --room-highlight: rgba(255, 255, 0, 0.3);
}
```

Console What's New

Highlights from the Chrome 100 update

View and edit @supports at-rules

The CSS @supports at-rules are now displayed and editable in the Styles pane.

Recorder panel improvements

Set custom selector attribute, support common selector attributes, rename recording and more.

DevTools

FINDING THE XPATH SELECT ELEMENT

The screenshot shows the govinfo.gov website with the Chrome DevTools 'Elements' panel open. The page title is 'Browse Congressional Hearings'. The 'Elements' panel shows the DOM tree with the following structure:

```
<div class="contentarea fw-tab-content" data-id="collection" data-role="content" data-cid="view1" style="display: block;">
  <div id="browsecontentId">
    <div role="alert" class="col-xs-12 alert alert-warning hide" id="browseErrors"></div>
    <div id="resultsId" aria-label="View The Results of your Browse here">
      <div tabindex="0" id="browsewidgetId">
        <div id="chrg">
          <div id="title-banner">...</div>
          <div class="collection-info">
            <div class="container">
              <div class="dashboard hidden-xs">...</div>
              <div class="row">
                <div class="col-xs-12">
                  <div class="hidden-xs">...</div>
                  <div class="visible-xs collapsed">...</div>
                  <div id="aboutDescription" class="collapse">
                    <p>...</p>
                    <div class="dashboard visible-xs">...</div>
                  </div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>
```

The XPath for the selected element is: `//div[@class='contentarea fw-tab-content' and @data-id='collection']`

The page content includes a description of hearings and a list of committees:

Browse Congressional Hearings

117th Congress (2021 - 2022)

- House Hearings
 - Committee on Agriculture
 - Committee on Appropriations
 - Committee on Armed Services

FINDING THE XPATH

COPY PREFERRED XPATH VERSION

The screenshot shows the govinfo website's "Congressional Hearings" page. The page includes a navigation bar with "govinfo", "Browse", "About", "Developers", "Features", "Help", and "Feedback". A "HELP" button is visible in the top right. The main content area features a definition of a hearing and a "Browse Congressional Hearings" section. Under "117th Congress (2021 - 2022)", there is a "House Hearings" list with categories like "Committee on Agriculture", "Committee on Appropriations", "Committee on Armed Services", and "Committee on Education and Labor". A specific hearing is highlighted: "Serial No. 117-1 (House Hearing) - Clearing the Path for Science-Based Strategies to Protect Workers from Covid-19 Infections".

Overlaid on the right is the Chrome DevTools interface. The "Elements" panel shows the DOM tree with the following structure:

```
<div class="contentarea fw-tab-content" data-id="collection" data-role="content" data-cid="view1" style="display: block;">
  <div id="browsecontentId">
    <div role="alert" class="col-xs-12 alert alert-warning hide" id="browseErrors"></div>
    <div id="resultsId" aria-label="View The Results of your Browse here">
      <div tabindex="0" id="browsewidgetId">
        <div id="chrg">
          <div id="title-banner"></div>
          <div class="collection-info">
            <div class="container">
              ::before
              <div class="dashboard hidden-xs"></div>
              <div class="row">
                ::before
                <div class="col-xs-12">
                  <div class="hidden-xs"></div>
                  <div class="visible-xs collapsed"></div>
                </div>
              </div>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
```

The "Styles" panel on the right shows the default Bootstrap styling for the container.

A context menu is open over the DOM tree, with the "Copy" option selected. The "Copy" submenu is visible, showing the following options:

- Copy element
- Copy outerHTML
- Copy selector
- Copy JS path
- Copy styles
- Copy XPath
- Copy full XPath

The "Copy XPath" option is highlighted, indicating the preferred method for finding the XPath.

SHORT XPATHS PROTECT AGAINST HTML CODING ERRORS

- ▶ Full XPath
`/html/body/div[10]/div[6]/div/div[2]/div/div[1]/div[2]/div/div[2]/div/div[3]/p`
- ▶ Short XPath
`//*[@id="aboutDescription"]/p`
- ▶ CSS selector (less flexible)
`#aboutDescription > p`

ADDING TO OUR SCRIPT: NAVIGATION + STORAGE

```
22  ## Taking screenshots
23  ## Very useful to check you're navigating right
24  my.session$takeScreenshot()
25
26  ## Navigating on the page
27  my.target= my.session$findElement(xpath='//*[@id="aboutDescription"]/p')
28
29  ## Extracting text (works reasonably well)
30  my.target$getText()
31
32
33  ## extracting the full HTML file
34  html = my.session$getSource()
35
36  page=read_html(scrape.page)
37  html=htmlParse(html, asText=T)
38  writeLines(text = html, con = 'data/my first page.html')
```

2. CREATE A LOCAL COPY

CAN YOU CAPTURE THIS DETAILS PAGE?

govinfo [Browse](#) [About](#) [Developers](#) [Features](#) [Help](#) [Feedback](#)

117th Congress (2021 - 2022)

House Hearings

- Committee on Agriculture
- Committee on Appropriations
- Committee on Armed Services
- Committee on Education and Labor

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

General. Thursday, March 11, 2021.

PDF TEXT **DETAILS** SHARE

check your result:
[url](#)

MORE XPATH: AXES

- ▶ parent or child
- ▶ ancestor or descendant
- ▶ preceding-sibling or following-sibling
- ▶ attribute

HTML PRIMER

DOCUMENT STRUCTURE - INDENTED

```
<table class="browse-node-table">
```

```
<tr id="this-is-the-1st-row">
```

```
<td colspan="2" id="this-is-the-only-cell">
```

```
S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
```

```
</td>
```

```
</tr>
```

```
<tr id="this-is-the-2nd-row">
```

```
<td id="this-is-one-cell">
```

```
<span class="results-line2">
```

```
Appropriation. Wednesday, April 24, 2013.
```

```
</span>
```

```
</td>
```

```
<td id="this-is-another-cell">
```

```
<a href="https://www.gpo.gov/link.to.PDF" target="_blank">
```

```
PDF
```

```
</a>
```

```
<a href="https://www.gpo.gov/link.to.text" target="_blank">
```

```
Text
```

XPATH RESOURCE

W3schools: Xpath tutorial

https://www.w3schools.com/xml/xpath_intro.asp

- ▶ Syntax
- ▶ Axes
- ▶ Special operators

3. MAKE CAPTURE COMPREHENSIVE

govinfo

[Browse](#)[About](#)[Developers](#)[Features](#)[Help](#)[Feedback](#)

Q SEARCH ▾

Committee on Education and Labor

Sort ↓_A^Z

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

PDFTEXTDETAILSSHARE

General. Thursday, March 11, 2021.

Serial No. 117-2 (House Hearing) - Rising to the Challenge: The Future of Higher Education Post Covid-19

PDFTEXTDETAILSSHARE

General. Wednesday, March 17, 2021.

Serial No. 117-3 (House Hearing) - Fighting for Fairness: Examining Legislation to Confront Workplace Discrimination

PDFTEXTDETAILSSHARE

General. Thursday, March 18, 2021.

Serial No. 117-4 (House Hearing) - Ending the Cycle: Examining Ways to Prevent Domestic Violence and Promote Healthy Communities

PDFTEXTDETAILSSHARE

General. Monday, March 22, 2021.

Serial No. 117-5 (House Hearing) - Lessons Learned: Charting the Path to Educational Equity

PDFTEXTDETAILSSHARE

COLLAPSE ALL

FOR LOOPS

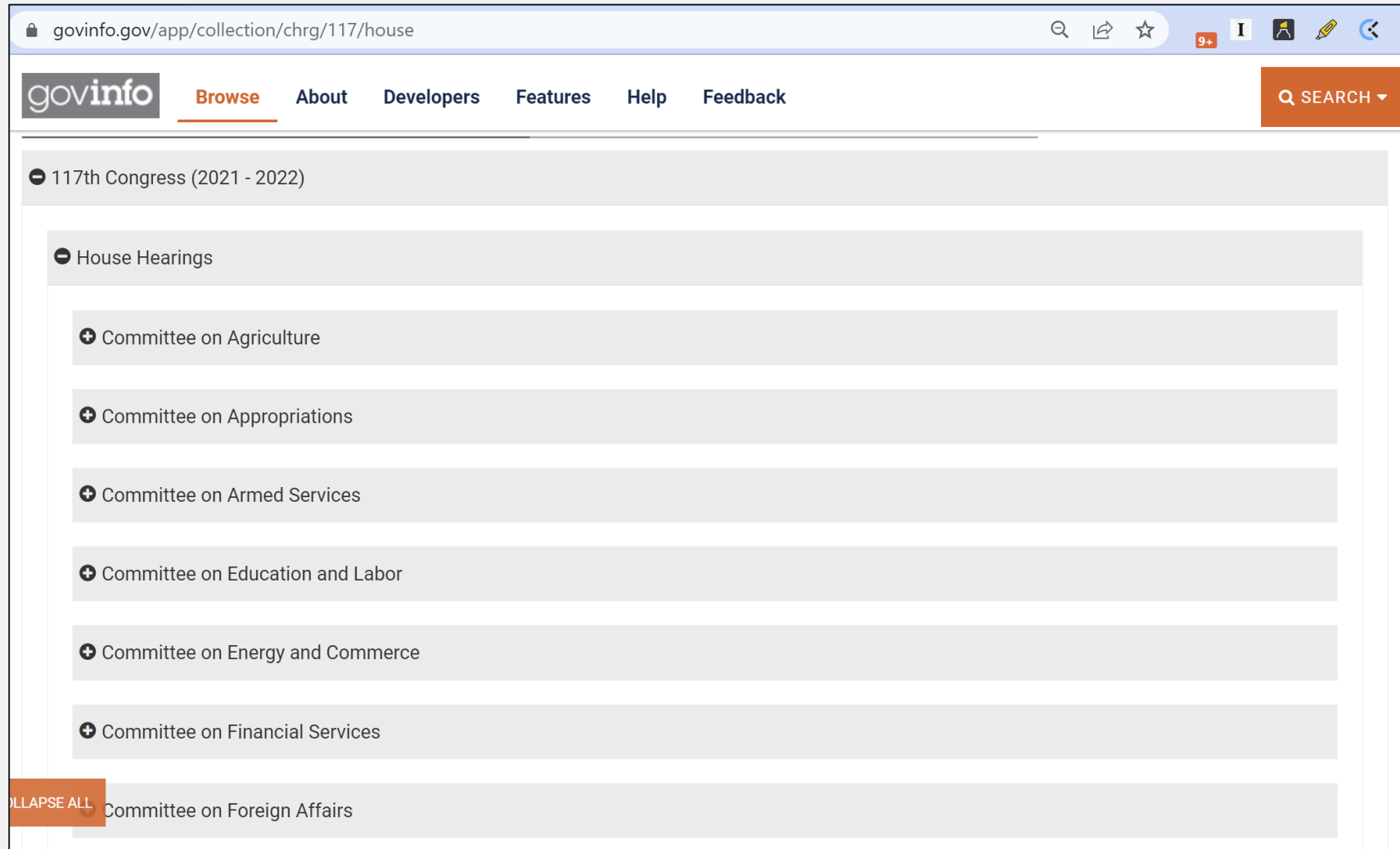
COLLECTING FROM THOUSANDS OF PAGES

- ▶ You have to repeat the same operation x times
- ▶ for loops avoid code repetition

```
for (x in 1:10) {  
  print(x)  
}
```

FOR LOOP EXERCISE:

PRINT THE NAMES OF ALL HOUSE COMMITTEES INTO YOUR R CONSOLE



PREPARING FOR NEXT WEEK

GETTING READY FOR GROUP WORK

1. Any project pitches?
2. Next week
 1. Parsing text using **Regular Expression**.
 2. Storing large data volumes in a **relational database design**.

THANK YOU!

END OF WEEK 1