# Lab 7: Rectangling and Webscraping

## Overview

This is a two part assignment. In the first part of the assignment you will practice rectangling on a dataset from the `repurrrsive` package. In the second part you will combine the `rvest` package along with functions and iteration to scrape data on foreign linked political action committees from the website open secrets.

## Rectangling

**Problem 1:** Load the `repurrrsive` package to get access to get access to the `got_chars` dataset. In section 23.4.2 of R4DS, there is code that extracts data from the `got_chars` list and converts it into a tibble with information on each character and a separate tibble which contains information on the titles held by each character. Perform similar operations to create separate tibbles containing the aliases, allegiances, books, and TV series of each Game of Thrones character.

## Webscraping Open Secrets

In the second part of this assignment we will scrape and work with data on foreign connected PACs that donate to US political campaigns. In the United States, only American citizens and green card holders can contribute to federal elections, but the American divisions of foreign companies can form political action committees (PACs) and collect contributions from their American employees.

First, we will get data foreign connected PAC contributions in the 2024 election cycle. Then, you will use a similar approach to get data such contributions from previous years so that we can examine trends over time.

You may benefit from using the Selector Gadget extension or similar tools.

In addition to `tidyverse`, you will need to install and load the packages `robotstxt` and `rvest`

**Problem 2:**

(a) Check that open secrets allows you to webscrape by running the `paths_allowed` function on the url `https://www.opensecrets.org`.

(b) Write a function called `scrape_pac()` that scrapes information from the Open Secrets webpage for foreign connected PAC contributions in a given year. The `url` for this data is https://www.opensecrets.org/political-action-committees-pacs/foreign-connected-pacs/2024. This function should take the url of the webpage as its only input and should output a data frame. The variables of this data-frame should be renamed so that they are in `snake_case` format (`lower_case_and_underscores_for_spaces`, see R4DS section 2.3). Use `str_squish()` to remove excess whitespace from the Country of Origin/Parent Company variables, and add a new column which records the year by extracting from the input url.

Hints (you may not need all of these):

- If you have trouble finding the right elements to search for using the selector gadget try looking for a table element.
- Use `read_html_live` instead of `read_html`
- To improve reliability, put a small pause after you use `read_html_live` (`Sys.sleep(1) should work`)
- Before your function returns, call `rm` to remove the object created by `read_html_live`. This prevents the creation of large numbers of browser sessions
- Sometimes your working function may fail inexplicably. Modify your function using `insistently` so that it retries multiple times.

(c) Test your function on the urls for 2024, 2022, and 2000, and show the first several rows of each of the outpus. Does the function seem to do what you expected it to do?

**Problem 3:**

(a) Construct a vector called `urls` that contains the URLs for each webpage that contains information on foreign-connected PAC contributions for a given year (combine `seq` and string functions). Using the `map_dfr` function from the `purrr` package, apply the `scrape_pac()` function over `urls` in a way that will result in a data frame called `pac_all` that contains the data for all of the years. I recommend the range 2000 to 2024.

(b) Clean this combined dataset by separating the country of origin from the parent company (use `separate_wider_delim` or another tool of your choice, you will need to be cautious with some special cases in this column) and by converting the strings in the `total`, `dems`, and `repubs` columns into numbers. Print out the top 10 rows over your dataset after completing these steps.

(c) Calculate the total contributions from PACs linked to Canada and Mexico each year and plot how these contributions change over time.

(d) Find the 5 countries who over the entire time period of the dataset have the greatest total contribution from affiliated PACs. Then calculate the total contribution for each of those countries for each year of the data and make a plot of it to visualize how the contributions have changed over time.