# Assignment 4: Web Scraping with R

This assignment demonstrates how to scrape and clean data from Wikipedia using R's rvest package. The tasks included: (1) scraping the foreign exchange reserves table by country, (2) modifying the script to scrape another table (currency composition), and (3) proposing a structured data plan for research applications.

## Task 1: Scraping Forex Reserves

The first table extracted from Wikipedia provides information on foreign exchange reserves by country. Below is the R code used to scrape and clean the table:

# Load packages  
library(tidyverse)  
library(rvest)  
library(stringr)  
  
# URL  
url <- 'https://en.wikipedia.org/wiki/List\_of\_countries\_by\_foreign-exchange\_reserves'  
wikiforreserve <- read\_html(url)  
  
# Extract first table  
foreignreserve <- wikiforreserve %>%  
 html\_nodes(xpath='//\*[@id="mw-content-text"]/div/table[1]') %>%  
 html\_table()  
  
fores <- foreignreserve[[1]]  
  
# Clean table  
fores\_clean <- fores[-c(1, 2), ]  
fores\_clean <- fores\_clean[, 1:6]  
names(fores\_clean) <- c("Rank", "Country", "Forexres", "Date", "Change", "Sources")  
fores\_clean$Date <- str\_split\_fixed(fores\_clean$Date, "\\[", n = 2)[, 1]  
fores\_clean$Date <- trimws(fores\_clean$Date)  
  
head(fores\_clean, 10)

## Task 2: Scraping Currency Composition

The second table provides the currency composition of reserves (COFER). The script was modified to scrape this table and then cleaned to focus on year, quarter, and USD values.

# Scrape second table  
other\_table <- wikiforreserve %>%  
 html\_nodes(xpath='//\*[@id="mw-content-text"]/div/table[2]') %>%  
 html\_table()  
  
other\_table\_df <- other\_table[[1]]  
  
# Clean table  
other\_clean <- other\_table\_df[-1, ]  
other\_clean <- other\_clean[, 1:4]  
names(other\_clean) <- c("Year", "Quarter", "USD\_billion", "EUR\_billion")  
  
# Fix column mix-up  
other\_clean <- other\_clean %>%  
 select(Year = Quarter, Quarter = USD\_billion, USD\_billion = EUR\_billion)  
  
head(other\_clean, 10)

## Task 3: Data Plan for Research

Web scraping with R and the rvest package offers a flexible way to collect real-time economic data from open sources such as Wikipedia, central banks, and international institutions. For research, a structured data plan is necessary:  
  
1. Identify reliable sources — official statistical agencies, IMF, World Bank, or trusted news outlets.  
2. Automate updates — schedule R scripts to scrape data at regular intervals.  
3. Clean and standardize variables — handle inconsistent formats and unify measurement units.  
4. Validate accuracy — compare scraped results with official datasets to avoid errors.  
5. Ethical considerations — ensure scraping complies with site policies (robots.txt) and cite sources properly.  
  
Following this plan allows researchers to build reproducible pipelines that integrate web data directly into economic and policy analyses, ensuring transparency and efficiency.