Data Collecting
– Web scraping from public websites

Jung PARK, PhD
Research Fellow in Data Science

Data Collecting

- Web scraping from public websites
  - How to collect data from websites
  - 10K reports from SEC.gov
  - EU Open Data
  - Open data Swiss
- Collecting data using API
  - Tweeter
  - Facebook
- Database in IMD library
  - Bloomberg
  - Thomson One
  - Datastream
  - Factiva: news media
- IoT (Internet of Things)
  - Smart building
  - Wearable devices
  - Web traffic

How to collect data from websites

- Traditional: Collect data by selecting texts or linked files manually

- New methods:
  - Accessing multiple files on a server using a code
  - Using API (Application Programming Interface)
  - Web scraping using a code

Public data sources (examples)

- Government data
  - 10K reports from SEC.gov
  - EU Open Data
  - Open data Swiss
  - Lond Data Store

- Public data
  - Wikipedia

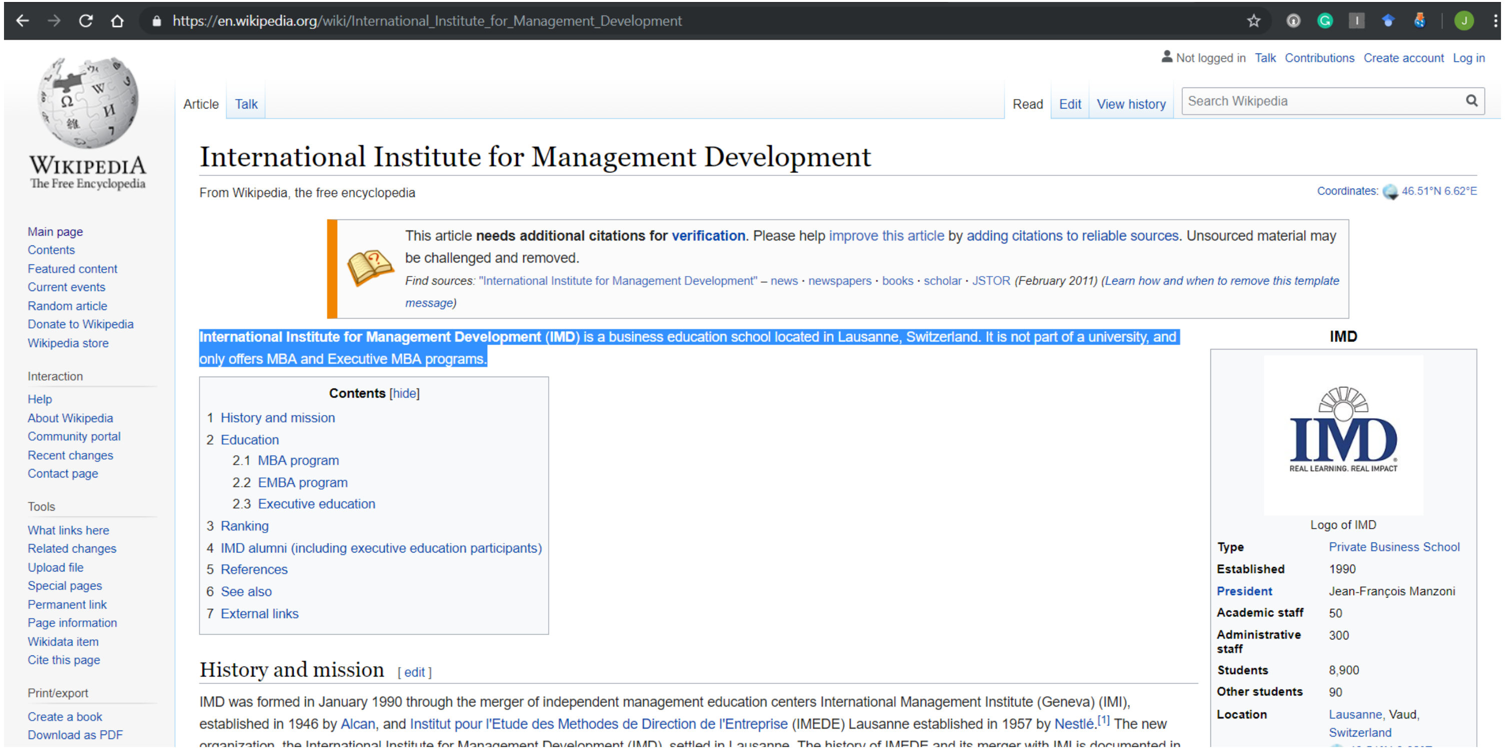
- Commercial data
  - Tweeter
  - Amazon

Data collecting approaches

- Level 1. Manual collection
- Level 2. Using a code to download multiple files
- Level 3. Using API (Application Programming Interface)
- Level 4. Using a code for Web scraping

Level 1: manual collection

- Goto the website; select texts; copy and paste using right mouse-click



https://en.wikipedia.org/wiki/International_Institute_for_Management_Development

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF

Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

International Institute for Management Development

From Wikipedia, the free encyclopedia

Coordinates: 46.51°N 6.62°E

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed.
Find sources: "International Institute for Management Development" – news · newspapers · books · scholar · JSTOR *(February 2011) (Learn how and when to remove this template message)*

International Institute for Management Development (IMD) is a business education school located in Lausanne, Switzerland. It is not part of a university, and only offers MBA and Executive MBA programs.

Contents [hide]

IMD

Logo of IMD

Type	Private Business School
Established	1990
President	Jean-François Manzoni
Academic staff	50
Administrative staff	300
Students	8,900
Other students	90
Location	Lausanne, Vaud, Switzerland

History and mission [edit]

IMD was formed in January 1990 through the merger of independent management education centers International Management Institute (Geneva) (IMI), established in 1946 by Alcan, and Institut pour l'Etude des Methodes de Direction de l'Entreprise (IMEDE) Lausanne established in 1957 by Nestlé.[1] The new organization, the International Institute for Management Development (IMD), settled in Lausanne. The history of IMEDE and its merger with IMI is documented in

Level 2: Using a code to download multiple files

```
# Construct a vector of 2 URLs
urls <- c("http://s3.amazonaws.com/assets.datacamp.com/production/course_1561/datasets/chickwts.csv",
"http://s3.amazonaws.com/assets.datacamp.com/production/course_3026/datasets/tsv_data.tsv")
for(url in urls){
        # Read a file in from the CSV URL
        csv_data <- read.csv(url)
        }
```

Level 3

```
# Construct a vector of 2 URLs
urls <- c("http://httpbin.org/status/404", "http://httpbin.org/status/301")
for(url in urls){
    # Send a GET request to url
    result <- GET(url, user_agent("my@email.address this is a test"))

    # Check request_result
    if(http_error(result)){
        warning('The request failed')
    } else {
        content(result)
    }

    # Delay for 5 seconds between requests
    Sys.sleep(5)
    }

    # Create list with nationality and country elementsquery_params
    <- list(nationality = "americans",      country = "antigua")     # Make
    parameter-based call to httpbin, with
    query_paramsparameter_response <-
    GET("https://httpbin.org/get", query = query_params)# Print
    parameter_responseparameter_response

    # Construct a directory-based API URL to `http://swapi.co/api`,#
    looking for person `1` in `people`directory_url <-
    paste("http://swapi.co/api", "people", "1", sep = '/')# Make a GET
    call with itresult <- GET(directory_url)
```

R code for collecting table information from Wikipedia

```
library(httr)
library(rvest)
library(xml2)

get_infobox <- function(title){
  base_url <- "https://en.wikipedia.org/w/api.php"

  # Change "Hadley Wickham" to title
  query_params <- list(action = "parse",
                       page = title,
                       format = "xml")

  resp <- GET(url = base_url, query = query_params)
  resp_xml <- content(resp)

  page_html <- read_html(xml_text(resp_xml))
  infobox_element <- html_node(x = page_html, css =".infobox")
  page_name <- html_node(x = infobox_element, css = ".fn")
  page_title <- html_text(page_name)

  wiki_table <- html_table(infobox_element)
  colnames(wiki_table) <- c("key", "value")
  cleaned_table <- subset(wiki_table, !wiki_table$key == "")
  name_df <- data.frame(key = "Full name", value = page_title)
  wiki_table <- rbind(name_df, cleaned_table)

  wiki_table
}

# Test get_infobox with "Hadley Wickham"
get_infobox(title = "Hadley Wickham")
```