



# **Data Collecting**

## **– Web scraping from public websites**

Jung PARK, PhD  
Research Fellow in Data Science

First uploaded 2019 July 30  
Last modified 2019 Aug 15

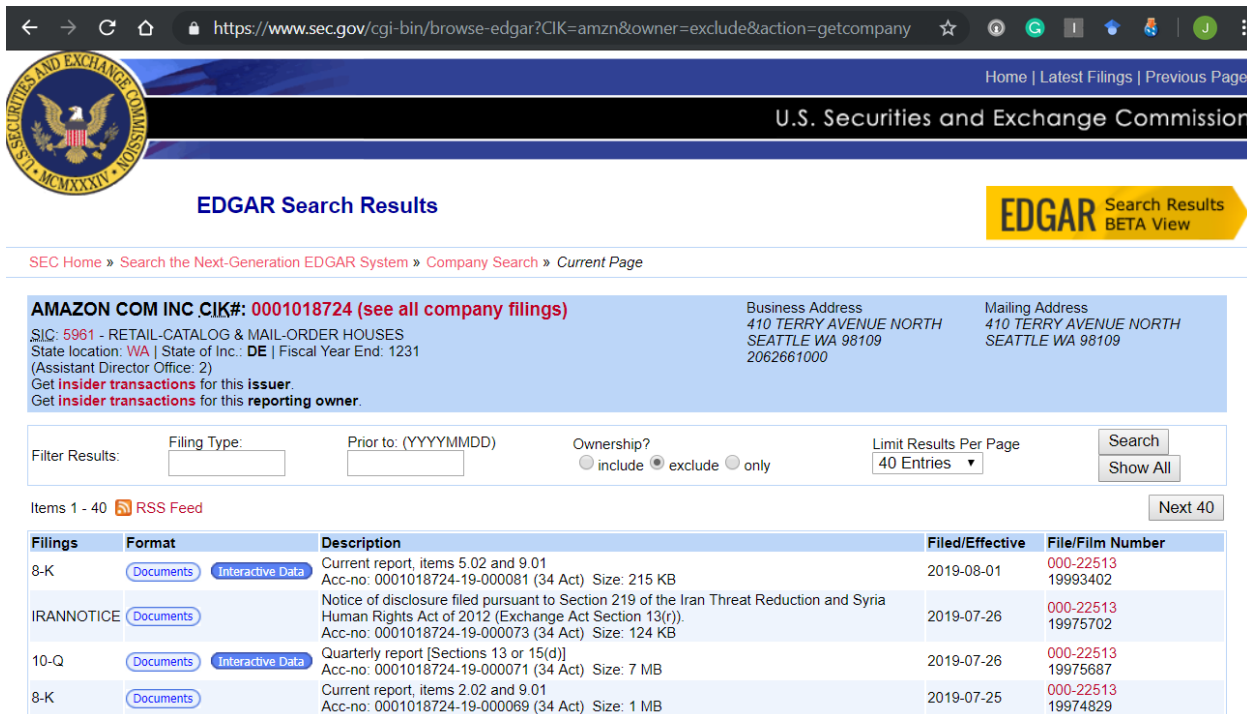
- Web scraping from public websites
  - How to collect data from websites
  - 10K reports from SEC.gov
  - EU Open Data
  - Open data Swiss
- Collecting data using API
  - Tweeter
  - Facebook
- Database in IMD library
  - Bloomberg
  - Thomson One
  - Datastream
  - Factiva: news media
- IoT (Internet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

- 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

## ex. Company filings from SEC.gov

- <https://www.sec.gov/edgar/searchedgar/companysearch.html>
- Mouse right-click and download individual files



The screenshot shows the SEC.gov Edgar Search Results page for Amazon.com Inc. (CIK# 0001018724). The page header includes the SEC logo and navigation links. The main content area displays the company's name, CIK number, and various filing details. Below this, there are filters for Filing Type, Prior to date, Ownership, and Limit Results Per Page. A table of filings is shown at the bottom, listing the filing type, format, description, and the date and number of the filing.

SEC Home » Search the Next-Generation EDGAR System » Company Search » Current Page

**AMAZON COM INC CIK#: 0001018724 (see all company filings)**

SIC: 5961 - RETAIL-CATALOG & MAIL-ORDER HOUSES  
State location: WA | State of Inc.: DE | Fiscal Year End: 1231  
(Assistant Director Office: 2)  
Get **insider transactions** for this **issuer**.  
Get **insider transactions** for this **reporting owner**.

Business Address  
410 TERRY AVENUE NORTH  
SEATTLE WA 98109  
2062661000

Mailing Address  
410 TERRY AVENUE NORTH  
SEATTLE WA 98109

Filter Results: Filing Type: Prior to: (YYYYMMDD) Ownership? ☐ include ☒ exclude ☐ only Limit Results Per Page 40 Entries Search Show All

Items 1 - 40 RSS Feed Next 40

Filings	Format	Description	Filed/Effective	File/Film Number
8-K	<a href="#">Documents</a> <a href="#">Interactive Data</a>	Current report, items 5.02 and 9.01 Acc-no: 0001018724-19-000081 (34 Act) Size: 215 KB	2019-08-01	000-22513 19993402
IRANNOTICE	<a href="#">Documents</a>	Notice of disclosure filed pursuant to Section 219 of the Iran Threat Reduction and Syria Human Rights Act of 2012 (Exchange Act Section 13(r)). Acc-no: 0001018724-19-000073 (34 Act) Size: 124 KB	2019-07-26	000-22513 19975702
10-Q	<a href="#">Documents</a> <a href="#">Interactive Data</a>	Quarterly report [Sections 13 or 15(d)] Acc-no: 0001018724-19-000071 (34 Act) Size: 7 MB	2019-07-26	000-22513 19975687
8-K	<a href="#">Documents</a>	Current report, items 2.02 and 9.01 Acc-no: 0001018724-19-000069 (34 Act) Size: 1 MB	2019-07-25	000-22513 19974829

- Easy to do; can check what we get immediately
- Boring and time-consuming
- We used to hire interns to do so

## Level 2: Using a code to download multiple files ex. Company filings from SEC.gov

SEC.gov explains how to access files from their server:

<https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm>

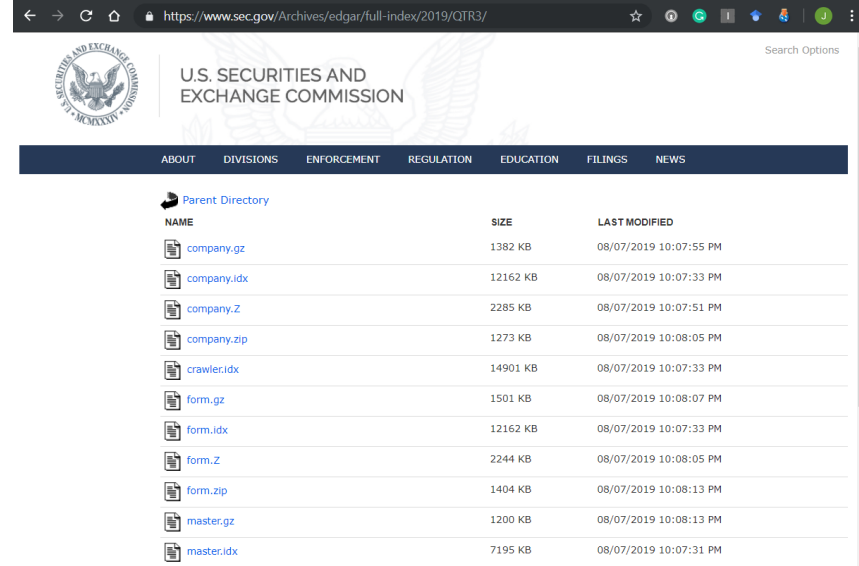
For example, we can use the file locations written in the master index provided by SEC.gov:

<https://www.sec.gov/Archives/edgar/full-index/2019/QTR1/master.idx>

master.idx

Description: Master Index of EDGAR Dissemination FeedLast Data Received: March 31, 2019  
Comments: webmaster@sec.gov Anonymous FTP: ftp://ftp.sec.gov/edgar/Cloud HTTP:  
<https://www.sec.gov/Archives/CIK/Company Name/Form Type/Date Filed/Filename>

1000045|NICHOLAS FINANCIAL INC|10-Q|2019-02-14|edgar/data/1000045/0001193125-19-039489.txt  
1000045|NICHOLAS FINANCIAL INC|4|2019-01-15|edgar/data/1000045/0001357521-19-000001.txt  
1000045|NICHOLAS FINANCIAL INC|4|2019-02-19|edgar/data/1000045/0001357521-19-000002.txt  
1000045|NICHOLAS FINANCIAL INC|4|2019-03-15|edgar/data/1000045/0001357521-19-000003.txt  
1000045|NICHOLAS FINANCIAL INC|8-K|2019-02-01|edgar/data/1000045/0001193125-19-024617.txt  
1000045|NICHOLAS FINANCIAL INC|SC 13G/A|2019-02-04|edgar/data/1000045/0001104659-19-005360.txt  
1000045|NICHOLAS FINANCIAL INC|SC 13G/A|2019-02-08|edgar/data/1000045/0001258897-19-001312.txt  
1000045|NICHOLAS FINANCIAL INC|SC 13G/A|2019-02-11|edgar/data/1000045/0001019056-19-000082.txt



U.S. SECURITIES AND EXCHANGE COMMISSION

NAME	SIZE	LAST MODIFIED
<a href="#">company.gz</a>	1382 KB	08/07/2019 10:07:55 PM
<a href="#">company.idx</a>	12162 KB	08/07/2019 10:07:33 PM
<a href="#">company.Z</a>	2285 KB	08/07/2019 10:07:51 PM
<a href="#">company.zip</a>	1273 KB	08/07/2019 10:08:05 PM
<a href="#">crawler.idx</a>	14901 KB	08/07/2019 10:07:33 PM
<a href="#">form.gz</a>	1501 KB	08/07/2019 10:08:07 PM
<a href="#">form.idx</a>	12162 KB	08/07/2019 10:07:33 PM
<a href="#">form.Z</a>	2244 KB	08/07/2019 10:08:05 PM
<a href="#">form.zip</a>	1404 KB	08/07/2019 10:08:13 PM
<a href="#">master.gz</a>	1200 KB	08/07/2019 10:08:13 PM
<a href="#">master.idx</a>	7195 KB	08/07/2019 10:07:31 PM

## Level 2: Using a code to download multiple files

ex. Company filings from SEC.gov

# python code

```
import requests
```

```
def download (url):  
    path = url.split("/")[-1]  
    r = requests.get(url)  
    open(path, 'wb').write(r.content)
```

```
urls =  
['https://www.sec.gov/Archives/edgar/data/320193/0000  
320193-19-000026.txt',  
'https://www.sec.gov/Archives/edgar/data/320193/00011  
93125-19-004664.txt']
```

```
for url in urls:  
    download (url)
```

Define a function for downloading  
Use the same file name obtained from url  
Get the file indicated by the url  
Write the content of the file indicated by  
the url as the name defined in path

Set url addresses (ex. from SEC.gov)

Call the download function repeatedly

## Level 3. Using API (Application Programming Interface)

- API is a standardized way for external users to collect data from a website's database
- Some websites like wikipedia provide API to encourages developing a third-party software providing additional services based on the company's product
- Also, this can avoid the abuse of data access using a risky method to the server
- reference: [https://www.mediawiki.org/wiki/API:Main\\_page](https://www.mediawiki.org/wiki/API:Main_page)

## Level 3. Using API (Application Programming Interface)

### ex. R code for collecting table information from Wikipedia

```
# R code
library(httr)
library(rvest)
library(xml2)

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

  query_params <- list(action = "parse",
                        page = title,
                        format = "xml")

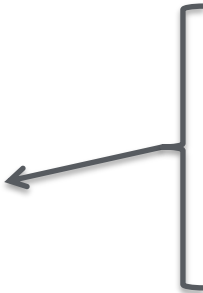
  resp <- GET(url = base_url, query = query_params)
  resp_xml <- content(resp)

  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
}

get_infobox(title = "International Institute for Management Development")
```



Call some packages  
Define a function get\_infobox  
We will use wikipedia api  
Send query parameters and receive data



Rearrange the data into a table form



Call the function for “IMD”



# Level 3. Using API (Application Programming Interface)

## ex. R code for collecting table information from Wikipedia

### Output

	key	value
1	Full name	IMD
2	Logo of IMD	Logo of IMD
3	Type	Private Business School
4	Established	1990
5	President	Jean-François Manzoni
6	Academic staff	50
7	Administrative staff	300
8	Students	8,900
9	Other students	90
10	Location	Lausanne, Vaud, Switzerland <span><span><span><span><span>46°31′N</span> <span>6°37′E</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span>&lt;U+FEFF&gt;﻿ / <span>&lt;U+...</span></span></span></span></span>
11	Campus	Urban
12	Website	<a href="http://www.imd.org/">http://www.imd.org/</a>
13	Alumni: 100,000	Alumni: 100,000

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

Printable version

Languages

Deutsch


Español

Français

### 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.

<b>Contents</b> <span>[hide]</span>
1 History and mission
2 Education <ul style="list-style-type: none"> <li>2.1 MBA program</li> <li>2.2 EMBA program</li> <li>2.3 Executive education</li> </ul>
3 Ranking
4 IMD alumni (including executive education participants)
5 References
6 See also
7 External links

### 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.<sup>[1]</sup> The new organization, the International Institute for Management Development (IMD), settled in Lausanne. The history of IMEDE and its merger

IMD



REAL LEARNING. REAL IMPACT

Logo of IMD

<b>Type</b>	Private Business School
<b>Established</b>	1990
<b>President</b>	Jean-François Manzoni
<b>Academic staff</b>	50
<b>Administrative staff</b>	300
<b>Students</b>	8,900
<b>Other students</b>	90
<b>Location</b>	Lausanne, Vaud, Switzerland <span><span><span><span><span>46°51′N</span> <span>6°62′E</span></span></span><span><span>﻿</span> / <span>﻿</span></span><span><span></span></span></span></span>
<b>Campus</b>	Urban
<b>Website</b>	<a href="http://www.imd.org/">http://www.imd.org/</a>
<b>Alumni: 100,000</b>	

## Level 4. Using a code for Web scraping

Even though there is no API provided, it is possible to extract texts from any websites using R and Python.

[to be updated]

## 10K and proxy statements from SEC.gov and context analysis

- Abraham Lu at IMD Global Board Center ([abraham.lu@imd.org](mailto:abraham.lu@imd.org))
- <https://www.sec.gov/edgar/searchedgar/accessing-edgar-data.htm>

## Webscribing codes

- <https://www.datacamp.com/courses/working-with-web-data-in-r>
- <https://www.tidytextmining.com/index.html>

- SEC.gov
- EU Open Data
- Open data Swiss
- London Data Store
- Wikipedia
- Federal Reserve Bank of St. Louis