Moving Closer Towards Web Data Integration

Manpreet Singh Juneja
Former FinTech Intern,CIB Global Research, JP Morgan Chase & Co.
manpreetsinghjuneja.becse17@pec.edu.in

Objective

1. To get a glimpse of Web Scraping and Web Data Integration
2. Learn to scrape web data using 3 different methods.

What to expect from this workshop?

1. Using WDI APIs
2. Data Cleaning
3. Data Extraction Logic Building
4. Using DevTools
5. Understanding XPaths
6. Using Selenium

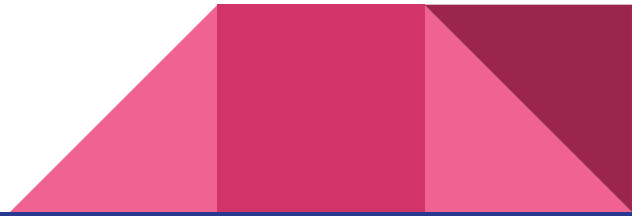
What is WDI?

WDI is an extension and specialization of data integration that views the web as a collection of heterogeneous databases.

Example Website: https://www.covid19india.org/

The Need for Web Data Integration

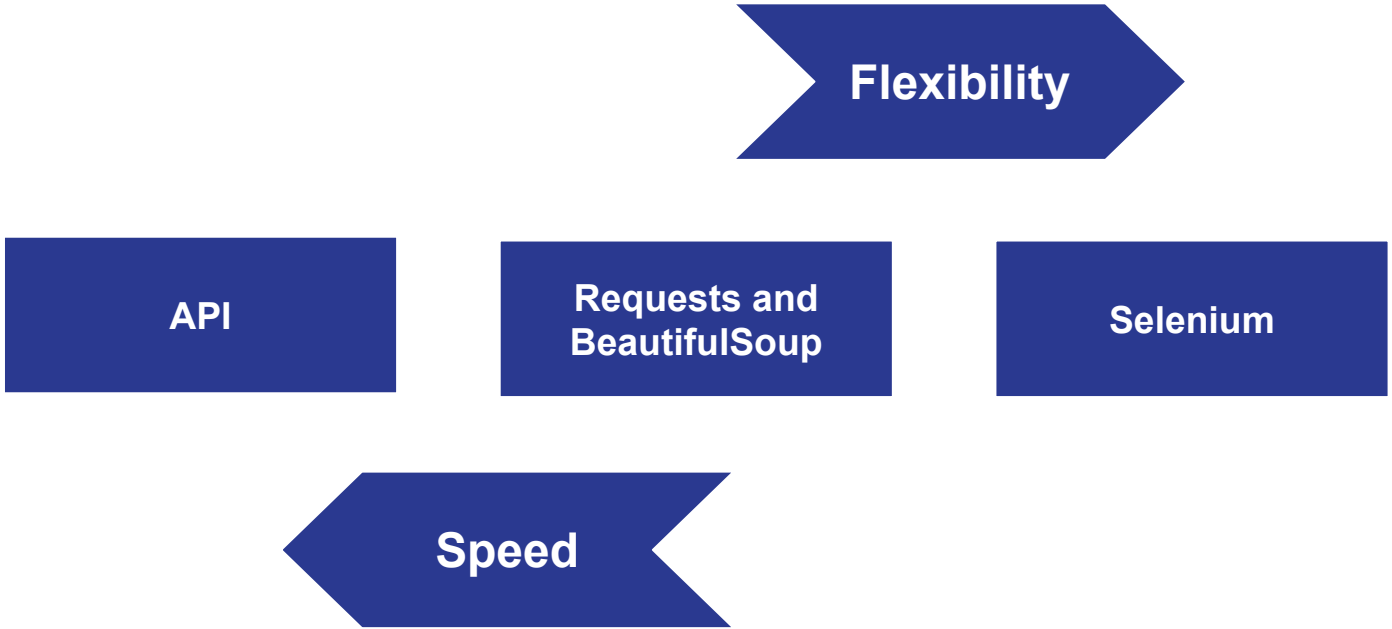
1. At the beginning of 2020, the digital universe was estimated to consist of 44 zettabytes of data.
2. Modern SMEs (Small and Medium Enterprises) have access to an unprecedented amount of data. Unfortunately, most of it goes unused in terms of analytics, meaning business owners are failing to capitalize on the data boom.
3. With AI-powered services growing in number, the greed for data is exploding.



The Stages of a WDI Life Cycle

1. **Identify**: Find the page, and where on the page, web data is located.
2. **Extract**: Collect either displayed or hidden content, even if it's behind a login or require user interactions.
3. **Prepare**: Cleanse and normalize web data using functions and formulas.
4. **Integrate**: Bring prepared web data into other business applications for faster insights.
5. **Consume**: Use graphs, charts, and other visualizations to see how web data changes over time.

Methods of Extracting Web Data



	Web Scraping	Web Data Integration
Data quality	Poor quality, unreliable and dirty data	High quality, reliable and clean data
Resource Needed	Resource intensive, requiring specialized skills	Little or no internal resources required
Legal Risk	High risk	Low or little risk
Integration	Stand alone files—needs to be integrated manually or with other tools	Integrated directly into business process and apps,analytics and AI platforms, and data warehouses—ready for consumption
Time to Value	Time consuming and difficult to maintain	Rapid onboarding and data delivery, and automatically maintained
Components	Scrape only, a subset of Extract	Extract, Prepare and Integrate
Types of data extracted	Displayed only	Displayed, Hidden and Derived
Scaling	Poor and fragile	Web scale

Source: import.io

Career Opportunities

Since this is a very upcoming field, not many companies offer a dedicated job profile for the same.

However, freelancing is quite popular and some startups do offer A dedicated role for the same.

1. https://www.upwork.com/freelance-jobs/web-scraping/
2. https://www.import.io/careers/
3. https://www.octoparse.com/blog/top-30-free-web-scraping-software?qu=

Ongoing Research

Dr. Manish Kamboj

References

1. https://www.import.io/post/web-data-integration-vs-web-scraping/
2. https://pupuweb.com/quality-control-web-data-enters-enterprise-mainstream/
3. https://realpython.com/learning-paths/python-web-scraping/