# WEB SCRAPING: A BIG DATA BUILDING TOOL AND ITS STATUS IN THE FINTECH SECTOR IN VIET NAM

**Nguyen Huong Anh**

*Institute of Post and Telecommunications Technology*

***Abstract:*** The amount of data in our lives is growing exponentially. With this surge, data analytics has become a hugely important part of the way organizations are run. And while data has many sources, its biggest repository is on the web. As the fields of big data analytics, artificial intelligence and machine learning grow, companies need data analysts who can scrape the web in increasingly sophisticated ways.

Web scraping (or data scraping) is a technique used to collect content and data from the internet. This data is usually saved in a local file so that it can be manipulated and analyzed as needed. If you've ever copied and pasted content from a website into an Excel spreadsheet, this is essentially what web scraping is, but on a very small scale. This paper will focus on various aspects of web scraping, beginning with the basic introduction and a brief discussion on various software's and tools for web scrapping. There are many benefits of web scraping in fintech sector such as market sentiment, compliance, risk mitigation, monitoring ratings changes, understanding customers, etc. The author had also found out some outstanding web scraping tools with some strengths and weaknesses and finally concluded with the web scraping's current status of application in the fintech sector in Vietnam.

***Keywords***: web scraping, big data, fintech, vietnam

## 1. INTRODUCTION

Presently the internet world is enormously enormous considering the web pages with huge quantity of explanatory substances obtainable with dissimilar designs such as text, graphical, audio-video, etc. which will focus on the contradiction in repossession of facts owing to the insignificance regarding the fact user is seeing. The data that is displayed by the websites can only be viewed using a web browser. They do not offer the functionality to save a copy of this data for personal use. The only option then is to manually copy and paste the data exhibited in the website at browser into the hard drive of our computer which is quite tiresome job. This is where web scraping comes into play.

Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. Commonly, web data is scrapped utilizing Hypertext Transfer Protocol (HTTP) or through a web browser. This is accomplished either manually by a user or automatically by a bot or web crawler.

To adapt to a variety of scenarios, current web scraping techniques have become customized from smaller ad hoc, human-aided procedures to the utilization of fully automated systems that are able to convert entire websites

Tác giả liên hệ: Nguyễn Hương Anh
Email: huonganh@ptit.edu.vn
Đến tòa soạn: 08/11/2022, chỉnh sửa: 15/12/2022, chấp nhận đăng: 06/01/2023

into well-organized data set. State-of-the-art web scraping tools are not only capable of parsing markup languages or JSON files but also integrating with computer visual analytics (Butler, 2007) and natural language processing to simulate how human users browse web content (Yi et al., 2003).

## 2. LITERATURE REVIEW

### Definition of Scraping

Scraping, also referred to as “web scraping” or “screen scraping”, is a method that allows third party companies (their developers) to access webpage data that users would normally have to log in to acquire. The collection of data is automated and the data is converted into formats of your choice, such as HTML, CSV, Excel, JSON, txt. (Zhao, 2017). Due to the fact that an enormous amount of heterogeneous data is constantly generated on the WWW, web scraping is widely acknowledged as an efficient and powerful technique for collecting big data (Mooney et al., 2015; Bar-Ilan, 2001).

Scraping a web page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page). Therefore, web crawling is a main component of web scraping, to fetch pages for later processing. Once fetched, extraction can take place. The content of a page may be parsed, searched and reformatted, and its data copied into a spreadsheet or loaded into a database. Web scrapers typically take something out of a page, to make use of it for another purpose somewhere else. An example would be finding and copying names and telephone numbers, companies and their URLs, or e-mail addresses to a list (contact scraping).

As well as contact scraping, web scraping is used as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup, and web data integration.

Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. As a result, specialized tools and software have been developed to facilitate the scraping of web pages. (Singrodia et al., 2019)

In summary, the process of web scraping primarily consists of 3 parts:

1. Parse through an HTML website

2. Extract the data needed

3. Store the data

### Why do we use scraping?

Web scraping, or Data scraping, is widely used across industries for the following reason listed below:

*Competitor Monitoring*

To keep tabs on competitors' strategies, businesses need to get fresh data from their competitors. This helps reveal insights into pricing, advertising, social media strategy, and many more (Persson, 2016).

For example, in the E-commerce industry, online store owners collect product information such as the sellers, images, and prices from websites like Amazon, Bestbuy, eBay, and AliExpress. This way, they can get first-hand market information and adjust their business strategy accordingly.

*Social media Sentiment Analysis*

Nowadays almost everyone owns at least one account on social media platforms like

Facebook, Twitter, Instagram, and YouTube (Datareportal, 2022). These platforms not only connect us with each other but also provide free space for us to express opinions publicly. People are so used to commenting online about things, such as a person, a product, a brand, and a campaign. Therefore, people collect comments and analyse their sentiments to help understand public opinions better.

In an article entitled Scraping Twitter and Sentiment Analysis using Python (Weldon, 2021), Ashley Weldon collected more than 10k tweets about Donald Trump and used Python to analyse the underlying sentiment. The result showed that the negative words in these tweets are way more diverse than the positive ones, which further indicated that people supporting him were generally less educated than people who disliked him.

Similarly, performing sentiment analysis allows businesses to know what their customers like or dislike about them, which helps them improve their product or customer service.

*Product Trend Monitoring*

In the business world, those who see the furthest ahead (and most accurately) are likely to win the competition. Product data empowers companies to predict the future of market trends more accurately (Persson, 2016).

In the case of the retailing industry, online fashion retailers scrape detailed product information to ensure an accurate estimate of demand. With a more thorough understanding of demand, there will be larger margins, faster-moving inventories, and smarter supply chains, which leads to higher income in the end.

*Monitoring MAP Compliance*

MAP compliance is a method for manufacturers to monitor retailers. In the retailing and manufacturing industries, manufacturers need to monitor retailers and make sure they comply with the lowest price. People need to keep track of the prices to stay competitive in the cut-throat market. With the help of web scraping, visiting all the websites and collecting the data are much more effective.

*Collect hotel & restaurant business information*

Another example of web scraping usage would be in the hospitality and tourism industry. Hotel consultants collect essential hotel information such as pricing, room types, amenities, locations from online travel agencies (Booking, TripAdvisor, Expedia, etc) to know about the general market price in a region. From there, they can improve the strategy for existing hotels or develop a strategy for starting new hotels. They also scrape hotel reviews and do sentiment analysis to know how the customers feel about their accommodation experience.

The same strategy applies to the dining industry. People collect restaurant information from Yelp, such as the names of the restaurants, categories, ratings, addresses, phone numbers, the price range to get an idea of the market they are targeting.

*News Monitoring*

Every minute, there are huge amounts of news generated global wide. Whether it is about a political scandal, a natural disaster, or a widespread disease, it’s not practical for anyone to read every piece of news from different sources. Web scraping makes it possible to extract news, announcement, and other relevant data from official and unofficial sources in a timely manner.

News monitoring helps notify important events happening all around the globe, and it assists governments in reacting to emergencies in no time. For instance, during the COVID-19 outbreak, the numbers of confirmed cases, suspected infections, and death tolls were constantly changing. Researchers can scrape the live & death statistics from China's government official website in real-time to further study and analyse the data. What's more, when countless reports and rumours were generated at the same time, the government was able to detect rumours among the facts quickly and clarify them, which reduces the possibility of unnecessary panic and even social chaos.

**Web scraping techniques**

*Traditional copy and paste*

Occasionally the human's manual examination and copy- and-paste method is the best and the workable web-scraping technology. But this is an error-prone, boring and tiresome technique when people need to scrap lots of datasets ("Web scraping," 2022).

*Text grapping and regular expression*

This is the simple and powerful approach to extract information from web pages. This technique based on the UNIX command or regular expression-matching facilities of programming language ("Web scraping," 2022).

*Hypertext Transfer Protocol (HTTP) Programming*

This technique used to extract data from static and dynamic web pages. Data can be retrieved by posting HTTP requests to the remote web server using socket programming ("Web scraping," 2022).

*Hyper Text Markup Language (HTML) Parsing*

Semi-structured data query languages, like XQuery and the Hyper Text Query Language (HTQL), can be used to parse HTML pages and to retrieve and transform page content ("Web scraping," 2022).

*Document Object Model (DOM) Parsing*

By embedding a full-fledged web browser, such as the Internet Explorer or the Mozilla browser control, programs can retrieve the dynamic content generated by client-side scripts. These browser controls also parse web pages into a DOM tree, based on which programs can retrieve parts of the pages ("Web scraping," 2022).

*Vertical aggregation platforms*

There are several companies that have developed vertical specific harvesting platforms. These platforms create and monitor a multitude of "bots" for specific verticals with no direct human involvement, and no work related to a specific target site. The preparation involves establishing the knowledge base for the entire vertical and then the platform creates the bots automatically. The platform's robustness is measured by the quality of the information it retrieves (usually number of fields) and its scalability (how quick it can scale up to hundreds or thousands of sites). This scalability is mostly used to target the Long Tail of sites that common aggregators find complicated or too labour-intensive to harvest content from ("Web scraping," 2022).

*Semantic annotation recognizing*

The pages being scraped may embrace metadata or semantic mark-ups and annotations, which can be used to locate specific data snippets. If the annotations are embedded in the pages, as Microformat does, this technique can be viewed as a special case of DOM parsing. In another case, the annotations, organized into a semantic layer, are stored and managed separately from the web pages, so the scrapers can retrieve data schema and instructions from this layer before scraping the pages ("Web scraping," 2022).

*Computer vision web-page analysers*

There are efforts using machine learning and computer vision that attempt to identify and extract information from web pages by interpreting pages visually as a human being might (“Web scraping,” 2022).

## 3. RESEARCH METHOD

In a time where vast amounts of data are being collected and archived by researchers all over the world, the practicality of utilizing existing data for research is becoming more prevalent (Andrews, Higgins, Andrews, Lalor, 2012; Schutt, 2011; Smith, 2008; Smith et al., 2011). Secondary data analysis is analysis of data that was collected by someone else for another primary purpose. The utilization of this existing data provides a viable option for researchers who may have limited time and resources. Secondary analysis is an empirical exercise that applies the same basic research principles as studies utilizing primary data and has steps to be followed just as any research method.

This paper uses secondary data analysis method. This study reviews and critiques existing conceptual frameworks proposed in both business and academic literature that dwell on web scraping, with the purpose of providing the information about web scraping in fintech sector in Vietnam.

This paper is organized as follows: a comprehensive review of relevant literature within industry and academia that offer information for web scraping and tools for web scraping. In the next section, the paper will examine the recent situation of web scraping in Fintech sector in Vietnam and will conclude by proposes some recommendations.

## 4. RESEARCH RESULTS

### Benefits of Web Scraping in Fintech

Financial decisions rely on good data. If you don’t have enough up-to-date information on a sector or a company, you risk making bad decisions. Those decisions need not be made by humans, but even AI needs data. So, what are some of the ways that web scraping can help businesses?

*Market sentiment*

The stock market is not entirely based on fundamentals. Sometimes stocks do surprising things, as the recent Reddit-fueled GameStop surges have shown. While it might have been challenging for any investor or hedge fund to be prescient enough to monitor Reddit for discussions related to GameStop, that kind of monitoring now seems prudent (just recently, Brooklyn ImmunoTherapeutics (BTX) surged, again probably because of Reddit).

Web scraping is ideal for watching websites for the shifts and tides in attitude that may herald erratic stock market behavior. The same goes for keeping an eye on social media to see what brands are being plugged by influencers. Traditional indicators, such as moving averages or the high-low index can also be efficiently and continuously monitored by web scraping. Plenty of websites offer these kinds of trackers, but any banking institution or fintech that wants to maintain its edge has good reason to roll their own and customize exactly what data they get and when they get it.

*Compliance*

At one time, fintech companies existed in a sort of internet Wild West, where they were, if not beyond the law, at least ignored by it. This is no longer the case, with fintech companies already being hit by fines of hundreds of thousands of dollars for not being compliant with financial regulations.

Luckily, governments and regulators are usually meticulous about publishing the rules online, which makes them available to web scraping tools. Apify recently helped a company to download 740,000 PDFs from over 100 government websites in less than 14 hours. While that was aimed at identifying economic and financial incentives for forest and landscape restoration in Latin America, it illustrates that government websites, designed to be easily accessible and public, can be web scraped. There's no need to be caught out by changes in the rules if you have tireless bots keeping an eye on updates and sending you automatic notifications.

*Risk mitigation*

Events in the real world can have devastating effects on companies and industries. Investors, insurance companies, and banks need to know when there are impending emergencies, such as freak weather or flooding, that could unexpectedly impact value. Web scraping can provide alerts on the first suggestion of unusual weather patterns from meteorological agencies or local news reports.

*Monitoring ratings changes*

Standard & Poor's (S&P), Moody's and Fitch are the three biggest rating agencies that regularly evaluate the credit worthiness of companies and governments. It may seem a small change, but if Moody's degrades a country by a single notch and it goes from 'A3' to 'Baa3', that will have a knock-on effect across the markets. It stands to reason that it would be a good idea to track these rating changes as they happen, without any delays in waiting for analysts to digest the information. Even better, there are other indices that appear to correlate with rating changes, such as the Corruption Perceptions Index. Gather enough information and you might find your tools predicting what Fitch is going to do. Even a margin of minutes could make all the difference.

*Investing wisely*

Due diligence is one side of checking out a company for potential investment, but online research can also provide useful information on both the company's news flow and the activities of its founders or staff. Social media profiles used to be considered hands-off for companies, but the rise of cancel culture means that increased attention is being paid to unacceptable online behavior and ignoring warning signs could be costly. Most investors would like to know where a company stands on touchy subjects, or even whether there are rumors of risky attitudes.

*Lead generation*

Just like any technology company, fintech companies need to market their solutions. While inbound and content marketing are going to be cornerstones of any fintech marketer, tracking down leads, especially when some of those leads might become huge, well-paying clients, is an approach that isn't going away. Reaching out to potential leads means that you need to know how to contact them and web scraping is a very efficient way to gather contact information. Rather than manually wading through websites to work out who your sales team should call, you can use a scraper to regularly crawl through hundreds of sites and extract the relevant details. If the person or any contact details change, your scraper will catch it and automatically flow the data right into your CRM system. Phone numbers, email addresses, social media profiles - anything relevant to keeping in touch with your customers can be scraped and stored.

*Understanding customers*

Scraping personal data is a minefield, especially after the European GDPR came into force in 2018. Recent EU regulations also restrict the kind of data fintechs can get through what is usually called screen scraping. Internet users might not mind sharing their personal details with companies like Facebook or Instagram, but they can get a little squeamish about the idea of other companies tracking them or keeping information about their habits in a database.

But fintechs thrive on serving customers better than old-fashioned financial institutions. While access to detailed financial data might be something to be extra careful about, fintechs don't have to make web scraping personal. Fintechs are newcomers to the financial scene so they need to keep up with how their brands and industry are seen across the web. Web scraping social media to see what customers are worried about, interested in, and how they talk about financial services is a non-invasive, ethical use of web scraping that pays dividends without the risk of negative press.

**Tools for Web scraping**

Over the last years, a set of companies, many of them start-ups, have realized that the market was demanding tools to extract, store, process and render data via APIs. The software industry moves fast and, in many directions, and a good web scraper can help in application development, Public Administration transparency, Big Data processes, data analysis, online marketing, digital reputation, information reuse or content comparers and aggregators, among other typical scenarios.

There are number of web scraping tools that available in market that can help user to scrape data from any website they want. Following are the list of some scraping tools. (Osmar, 2015)

*Import.io*

Indubitably, this is one of the reference tools in the market. It may be used in four different ways. The first one (named Magic, that can be classified as basic) is the access to import.io in order to type the address of the web site on which we want to perform scraping. The result is shown in an attractive visual tabular format. The main con is that the process is not configurable at all.

The second way of use is named Extractor. It is the most common usage of import.io technology: download, install and execute in your own computer. This tool is a customized web browser available for Windows, OS X and Linux. This way requires some previous skills using software tools and some time to learn how to use the tool. However, “picking” is offered in a quite reasonable manner -although open to improvement, at the same time. “Picking” is performed by clicking on the parts of the scraped web site that we want to extract, in a simple and visual way. This is a feature that any web scraping tools must include these days.

Once that queries have been created, output formats are only two: JSON and CSV. Queries may also be combined in order to page results (“Bulk Extract”) or aggregate them (“URLs from another API”). It is also relevant to note that users will have a RESTful API EndPoint to access the data source -which is a mandatory feature in any relevant complete scraping tool nowadays.

The import.io application requires a simple user registration process. With a username and password, the application can be used for free, with a set of basic features that may be sufficient for small developer teams without complex scraping requirements. For companies or professionals demanding more flexibility and backend resources, contact with import.io sales team is required.

The third and fourth ways of use provide more value to the tool. They are the Crawler and the Connector, respectively. The Crawler tries to follow all the links in the document indicated via its URL and it allows information extraction based on the picking process carried out in the initial document. In the tests carried out to write this article, we have not managed to finish all this process, as it seems to keep working all the time without producing any results. The Connector permits to record a browsing script to reach the web document from which to extract information. This approach is very interesting, for instance, if the data to be scraped are the result of a search.

In summary, import.io is a tool with interesting functionality free to use, with a high level of maturity, an attractive and modern graphical user interface, supporting cloud storage of the queries, which demands local installation to take full advantage of its features.

*Table 1. Strenghts vs Weaknesses of import.io*

| **Strengths** | **Weaknesses** |
|---|---|
| Visual Interface | Desktop installation |
| Blog and Documentation | Limited amount of output document formats |
| Allows pagination, aggregation, crawling and script recording | Learning curve |

Source: Provided by the author

*Kimonolabs*

Kimono Labs is another key player in the field of web scraping. It uses a strategy similar to import.io, using a web browser. In their case, they offer a Chrome extension, rather than embedding a web browser in a native desktop application. Therefore, the first step to use this tool is to install Google Chrome and then this extension. Registration is optional. It is a quick and simple process that is recommended.

After installation and registration, we can use Chrome to reach a web document with interesting information. In this moment, we click the icon installed by the Kimono extension and the picking process starts. This process provides help to the user at first execution with a visual and attractive format. Its graphical user interface is really polished and the tool results very friendly.

To start working with Kimono, the user must create a “ball” in the upper end of the screen. By default, a ball is already created. The various balls created are shown with a different colour. Afterwards, sections of the document may be selected to extract data an, subsequently, they are highlighted with the colour of the associated ball. At the same time, the ball shows the number of elements that match the selection. Balls may be used to select different zones of the same document, although their purpose is to refer zones with certain semantic consistency.

For instance, in the web site of a newspaper, we might create a ball named “Title” and then select a headline in a web page. Then, the tool highlights one or two additional headlines as interesting. After selecting a new headline, the tool starts highlighting another 20 interesting elements in the web page. We may notice that there are some headlines which are not highlighted by the tool. We select one of them and now over 60 headlines are highlighted. This process may be repeated until the tool highlights all the headlines after selecting a small amount of them. This process is known as “selector overload” and is available in several scraping tools.

Once that all the headlines have been selected, we can try doing the same with the opening paragraph of each news item: create a ball, click on the text area of an opening paragraph, then another one, and go on with the process until having all the desired information ready for extraction.

Although the idea is really good, our tests have found that the process is somehow bothersome. Sometimes, box highlighting in web pages does not work well with, for instance, problems in texts which are links at the same time.

Once that the picking process is finished by clicking on the "Done" button, we can name our new API and parameterize its temporal programming. With this, the system may execute the API every 15 minutes, hour, day, etc. and store the results in the cloud. Whenever the user calls the API by accessing the associated URL, Kimono does not access the target site but returns the most recent data stored in its cache. This caching mechanism is highly useful but not exclusive of Kimono.

The query management console, available at www.kimonolabs.com, provides access to the newly created API and various controls and panels to read data and configure how they are obtained. This includes an interesting feature, which are email alerts to be received when data change in the target site.

There is an additional interesting option named "Mobile App" that integrates the content of the created API in a view resembling a mobile application, allowing some styling configuration. However, the view generated by this option is a web document accessible by the URL announced, aimed to be rendered in a mobile browser. Unluckily, the name of the option misleads users and does not generate a mobile application to be published in any mobile application store. Still, it may be a useful option for rapid prototyping.

The console menu also offers the "Combine APIs" option. Initially, it may look like an aggregator, assembling the data obtained from several heterogeneous APIs in a single API. Nevertheless, help information in this option indicates that the aggregated APIs must have the same exact name of data collections. The conclusion is that this option is useful to paginate information, but not to aggregate.

In summary, kimono is a free tool, with a high level of maturity, a very good graphical user interface, providing cloud storage for queries, requiring Chrome browser and their extension - both of them installed locally.

*Table 2. Strenghts vs Weaknesses of Kimono Labs*

| **Strengths** | **Weaknesses** |
|---|---|
| Visual interface | Chrome browser dependency |
| Documentation | Does now allow aggregation |
| Picker | Weak mobile app option |

Source: Provided by the author

*myTrama*

myTrama is a new web crawling tool positioned as a clear competitor to those previously commented. It is a purely SaaS service, thus avoiding the need for users to install any software nor to depend on a specific web browser. myTrama works on Chrome, Firefox, Internet Explorer and Safari. It is available at https://www.mytrama.com.

A general analysis of this tool suggests that myTrama takes the best ideas of import.io and kimono. It presents information in a graphical user interface, perhaps not so good but more compact and with the look and feel of a project management tool. Some of the features which seem more interesting in this tool are commented below:

- Main view is organized in a way similar to an email client, with 3 vertical zones: 1) folders,

2) queries, and 3) query detail. It is efficient and friendly.

- Besides JSON, XML and CSV, the classical structured formats for B2B integrations, it adds PDF for quick reporting and sends results in an easily viewable and printable format.

- It includes a query language named Trama-WQL (quite similar to SQL), which is simple to use while powerful. It is useful when visual picking is not sufficient, providing a programmatic manner to define the picking process. Documentation of this language is available in the tool as a menu option.

- The "Audit" menu option gives access to a compact control panel with information about the requests currently being made to each of the APIs (EndPoints).

- The picker is completely integrated. It is not necessary any type of additional software. It is similar to the approach used in kimono, although it uses "boxes" instead of "balls". A subtle differentiation is that a magic wand replaces the default mouse pointer when picking is available. In addition, the picking process may be stopped by right-clicking on the area being picked.

- myTrama permits grouping boxes within boxes, although only one level of grouping and only a group with query are allowed. This is a very useful feature in order to have results properly grouped. Hopefully, the development team will improve this feature soon to provide users with more flexibility.

- Query configuration allows update frequency with a granularity of minutes, from 0 to 9999999. Zero means real time (this is, accessing the target site upon each request to the EndPoint). For any other value, information is obtained from the cache -as in kimono.

- APIs may be programmed using parameters sent via GET and POST requests. Unfortunately, the dev team has not published sufficient documentation related to this feature. For example, it is possible to use the URL of an API and overwrite the FROM parameter (the URL referencing the target document) in real time. It is also possible to pass parameters via GET and POST in the same API. Additionally, there is a service that allows the execution of a Trama-WQL sentence without any query created in the tool. As these are not very well documented features, the best choice is to contact the people at Vitesia.

- Paging queries and aggregation of heterogeneous queries are supported in a fairly simple and comfortable way.

- For those preferring the browser extension way of scraping, a Chrome extension is also available. This mechanism allows users to browse sites and start the scraping process by clicking on the button installed by the extension. This plugin is not yet published but can be requested to Vitesia.

- PDF is not only a format available as output, but also as input. Therefore, a URL may reference only HTML documents but also PDF. For instance, users will be able to extract information from PDFs and generate JSON documents that feed a database for later information analysis. The business hypothesis to support this is based on the evidence initially commented at the introduction of this article that stated that 70% of the content published in the Internet is contained in PDF documents. Vitesia consider that this may be a differentiating feature between myTrama and their competitors.

- APIs preserve session. This allows chaining calls to queries in myTrama and fulfil business processes, such as searches, step-based forms (wizards) or access information available behind a login mechanism.

- It is available in two languages: English and Spanish.

- Access to this platform is based on invitation. Users remain active for 30 days. Later contact with the dev team is required in order to move to a stable user.

Among all the tools analysed, myTrama seems to be the most complete and compact, although its user interface is one step behind kimono and import.io. For users with software development skills, myTrama seems to be the best choice -although requiring direct contact with Vitesia.

In summary, myTrama is a tool solely offered as a SaaS service, very complete to carry our scraping processes, with cloud storage and that may be operated with any web browser. Its major weakness is the lack of documentation of many differentiating issues relevant to developers interested in taking advantage of scraping processes.

*Table 3. Strenghts vs Weaknesses of myTrama*

| Strengths | Weaknesses |
|---|---|
| The Trama-WQL language | Limited documentation |
| Dashboards | Free license only for 30 days |
| Picker | More oriented to developers |

Source: Provided by the author

## Web Scraping status in Vietnam

### *Status of Fintech sector in Viet Nam*

In recent years, Asia has developed into one of the leading fintech regions, having recorded the highest fintech revenue worldwide. Vietnam is among the emerging fintech markets in this region. The country has an increasingly connected population, which creates a premise for financial technologies to develop. The fast-growing FinTech market in Vietnam holds great market potential for technology companies that support digital banking, digital payments, block chain, and cryptography. Vietnam is currently home to more than 130 FinTech startups that cater to numerous clients and cover a broad range of services such as digital payments, alternative finance, wealth management and blockchain, among others (Mordor Intelligence, 2022).

In spite of the negative socio-economic consequences of Covid-19, Viet Nam market still experienced a rapidly rising expansion of fintech sector. For example, the COVID-19 pandemic has accelerated the development of microinsurance solutions between fintech businesses and insurance companies. For instance, PVI and MoMo introduced the Corona++ nano insurance product to help clients of LMI (Lender Mortgage Insurance) who were at risk.

### *Status of Scraping in Fintech in Viet Nam*

The remarkable rise of Viet Nam Fintech’s industry has prompted entrepreneurs to constantly reach out to potential customers and be innovative in order to gain an edge against their rivals in such a competitive market. In developed market in foreign countries, the application of scraping has been proven extremely beneficial for fintech companies in collecting customers’ data, automating the extraction and aggregation of financial data, facilitating equity search, and enabling data-driven market predictions and branching out to provide additional products and services based on the data collected.

However, the application of scraping in Viet Nam is not entirely widespread and universally accepted due to the following reasons:

*Lack of legal framewor*

Viet Nam Law does not have concrete requirements regarding the usage of scraping and confidentiality of customers as of the moment; therefore, the legality in using scraping to collect data is quite dubious. As a result, the responsibility to protect customers' confidentiality falls on the owners of websites in which customers' information resides. As such, a number of companies and customers advocate for the restriction of scraping, or outright ban the usage of scraping for fear of their customers' data being misused for illegal purposes.

*Unfairness in competition*

Shopee - an online shopping platform, ban the use of scraping due to its usage being listed as "illegal data collection for the purpose of unfair competition". Similarly, there are sentiments among fintech companies in Viet Nam that application of scraping is essential exploiting a loophole in legality to compete unfairly in the market.

*Scraping vs API*

API (Application Programming Interface) is a commonly used software intermediary that facilitates communication and sharing of data among parties, provided the owners of data approve (Benslimane et al., 2008). API usage is legal in Viet Nam and therefore, has been adopted in most of financial institutions in Viet Nam in order to share customers' data in a secure and legal method. Moreover, even traditional banking sector in Viet Nam has also integrated API into their system. As a result, an aggressive and legally dubious tool such as scraping may find it hard to gain a foothold in the market when a safer and more proven alternative like API has already existed.

## 4. RECOMMENDATIONS

We are eager to keep expanding and refining the scope of research on scraping in order to deepen our understanding of scraping and its effect on Viet Nam market. Our current belief is that scraping has proven to be an effective tool in data collection in Viet Nam fintech sector. However, this is not prevalent in all companies in the market. A failed adoption can be attributed to a lack of understanding, insufficient executive attention, or inadequate financial investment. The challenges that arise from being a new and disruptive technology have yet to be fully specified; how best to meet those challenges is still to be discovered.

There are far too many other forces are in play, aside from the factors from the incumbents in the market. One of these essential forces is the government. From our research on scraping and its effect, we propose the following policies as follows:

**Establish policies to assist businesses in developing data service**: Assistance from the government not only provides financial incentives for businesses to develop data service, but also ensures that businesses feel safe and secure when adopting new technologies by issuing official approval and guidance for the service.

**Develop proper legal framework to protect customers' confidentiality**: Rising number of attack on customers' personal information in Viet Nam reflect the current problem of information security, especially in financial sector. As we have already mentioned that the responsibility to protect customers' confidentiality falls on the owners of websites in which customers' information resides due to the lack of legal framework, it is imperative that the government join forces with businesses in protecting customers' data, specifically by providing guidance in data protection and establishing legal protocols in dealing with data theft crime. Businesses and customers will be more accustomed to and be more comfortable with the use of scraping if they can be certain that the government is on their sides and will protect their interests.

## 5. CONCLUSION

Web scraping is a recognizable phrase which has expanded significance owing to the requirement of "free" data accumulated in web pages. Many professionals and researchers need the data in order to process it, analyse it and extract meaningful results. On the other hand, people dealing with B2B use cases need to access data from multiple sources to integrate it in new applications that provide added value and innovation.

Throughout this paper, the author has reviewed the various aspects of Web Scraping. Starting with the definition, techniques and tools for web scrapping, we have seen the benefits of Web Scraping in Fintech sector and finally viewed the status of web scrapping in Vietnam.

## ACKNOWLEDGEMENTS

The research **(MS: -HV-2022/TCKT)** was funded by the Post and Telecommunications Institute of Technology PTIT. Thank you for the generous donation I receive through PTIT and I am so grateful!

## REFERENCES

[1] Bar-Ilan, J. (2001). Data collection methods on the web for informetric purposes – A review and analysis. Scientometrics, 50(1), 7–32.

[2] Benslimane, Djamal, Schahram D., and Amit S. (2008). Services Mashups: The New Generation of Web Applications, IEEE Internet Computing, vol. 12, no. 5. Institute of Electrical and Electronics Engineers. pp. 13–15. .

[3] Butler, J. (2007). Visual web page analytics.

[4] Datareportal. (2022). Digital 2022 Global Overview Report. Retrieved September 20, 2022, from https://datareportal.com/reports/digital-2022-global-overview-report.

[5] Mooney, S. J., Westreich, D. J., & El-Sayed, A. M. (2015). Epidemiology in the era of big data. Epidemiology, 26(3), 390.

[6] Mordor Intelligence (2022). Vietnam Fintech Market Size, Share, Growth 2022-27. Mordor Intelligence.

[7] Osmar, C. (2015). Web Scraping: Applications and Tools. European Public Sector Information Platform Topic Report No. 2015 / 10. Retrieved September 19, 2022.

[8] Persson, J. G. (2016). Current trends in product development. Procedia CIRP, 50, 378–383. https://doi.org/10.1016/j.procir.2016.05.088

[9] Singrodia, V., Mitra, A., & Paul, S. (2019). A review on web scrapping and its applications. 2019 International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/iccci.2019.8821809

[10] Statista (2022). Digital payment users in Vietnam 2017-2025, by segment.

[11] Weldon, A. (2021). Scraping Twitter and Sentiment Analysis using Python [web log]. Retrieved September 25, 2022, from https://www.octoparse.com/blog/text-mining-and-sentiment-analysis-using-python#.

[12] Yi, J., Nasukawa, T., Bunescu, R., & Niblack, W. (2003). Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. Data Mining, 2003. ICDM 2003. Third IEEE International Conference on, IEEE. Melbourne, Florida, USA.

[13] Zhao, B. (2017). Web scraping. Encyclopedia of Big Data, 1–3. https://doi.org/10.1007/978-3-319-32001-4_483-1

## WEB SCRAPING: CÔNG CỤ XÂY DỰNG DỮ LIỆU LỚN VÀ THỰC TRẠNG TẠI VIỆT NAM

***Tóm tắt***: Lượng dữ liệu trong cuộc sống của chúng ta đang tăng lên theo cấp số nhân. Với sự gia tăng này, phân tích dữ liệu đã trở thành một phần cực kỳ quan trọng trong cách điều hành tổ chức. Và trong khi dữ liệu có nhiều nguồn, thì kho lưu trữ lớn nhất của nó là trên web. Khi các lĩnh vực phân tích dữ liệu lớn, trí tuệ nhân tạo và máy học phát triển, các công ty cần những nhà phân tích dữ liệu có thể quét web theo những cách ngày càng tinh vi.

Web scraping (hoặc quét dữ liệu) là một kỹ thuật được sử dụng để thu thập nội dung và dữ liệu từ internet. Dữ liệu này thường được lưu trong một tệp cục bộ để có thể thao tác và phân tích khi cần thiết. Bài báo giới thiệu về web scraping là một công cụ quét web để xây dựng dữ liệu lớn. Nghiên cứu tập trung vào việc làm rõ so sánh về các kỹ thuật web scraping và các công cụ phổ biến để web scraping. Ngoài ra, bài báo cũng sẽ cung cấp thông tin về web scraping ở Việt Nam.

***Từ khóa:*** web scraping, quét dữ liệu, công nghệ tài chính, dữ liệu lớn, Việt Nam

**Nguyen Huong Anh** graduated with a Bachelor of Finance, National Economics University, Vietnam in 2018; Master's degree in International Business, UK in 2020. She is now a Lecturer in the Faculty of Finance and Accounting at the Post and Telecommunications Institute of Technology (PTIT), Hanoi, Vietnam.

Research Interests: Financial Technology, Big Data, Corporate Finance