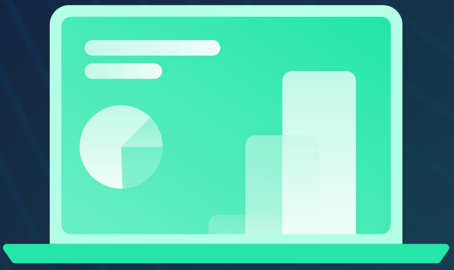
oxylabs®

Building a Real-Time Online Media Monitoring Infrastructure



Introduction

Tracking media coverage and customer sentiment is the backbone of successful brand monitoring and public relations. Overseeing brand mentions and other topics of interest online can significantly boost the understanding of the overall discourse surrounding the brand and the industry.

Not only does it help gain insight into the volume of mentions or estimated brand reach, but into the sentiment behind customer feedback and the context your brand is associated with as well. All this, in turn, contributes to better strategic planning and business decision-making.

To grasp the entire picture, however, and react immediately in case of emergency, you need to comb through myriads of web pages in real time. Performed without proper tools, it can be daunting and resource-inefficient.

Before embarking on this mission, you should put thought into building a real-time media monitoring infrastructure. It should be able to bring fruitful results and function in the long term, regardless of the project's scope and requirements.

Online media monitoring explained

Online media monitoring is the activity of tracking and collecting publicly available pieces of information about your or your clients' subjects of interest across multiple digital resources.

How does it work?

Companies that provide online media monitoring services offer their clients specialized platforms where they can type in a required keyword or phrase, such as a brand's name, and access all such mentions on the internet.

To create these media monitoring systems, companies need to use dedicated web crawling and scraping software that scouts for public data all over the web and retrieves it based on the predetermined parameters.

Potential

Online media monitoring opens up quite a few exciting opportunities for businesses. By performing regular and methodological online media monitoring, companies can benefit from it in three primary ways.

1. Measuring brand awareness

Brand awareness refers to how familiar target audiences are with a brand and its products. Normally, brands put tremendous effort into making themselves recognizable, as this factor is the driving force behind trust and, eventually, sales generation.

All this hard work may go to waste, however, if companies don't have certain processes set up to monitor and measure brand awareness. Simply monitoring brand mentions online can provide valuable insight into your coverage and popularity. The monitoring results will contain a specific number of mentions over the web with a list of all web pages and resources where your brand has been spotted.

2. Reputation management

Online media monitoring involves gathering information about the context in which your brand was mentioned. It means you can keep track of the sentiments associated with your brand and thus manage your reputation.

Real-time data collection allows you to promptly prepare action plans in case of any unfavorable mentions. Depending on the severity of your situation, you might want to contact the person(s) directly or make a public statement regarding any accusations. Real-time data gathering allows you to initiate these actions in a timely manner.

3. Competitor monitoring

Not only can you oversee your own or your client's brand mentions, but the industry and competition at large as well. By collecting competitor data and comparing it to your own metrics, you can learn how your brand performs and shape the business strategy more efficiently.

For example, you can keep an eye on how frequently and in what context your competitors are mentioned on media platforms central to your business and then compare it to your brand's media coverage. If the odds are not in your favor, you should figure out how to improve positions and upstage the competition.

Challenges

As is usually the case, nothing good comes easy, and online media monitoring is no exception. Just as rewarding as it is, it contains certain challenges.

1. Geo-restricted content

If your brand is expansive enough and penetrates global markets, you'll most likely want to monitor it in many different corners of the world at once. However, some content rests under geo-restrictions or, at least, may be displayed differently depending on the location you're trying to access it from. Thus, monitoring your brand and the topics of interest may require advanced resources to unlock geo-restricted data.

2. Security measures of data sources

Some websites can be especially protective of their content and implement various security measures to prevent their data from being scraped. Constant IP blocks, CAPTCHAs, and other measures can be severe obstacles on your way to the desired data. As such, it's important to find a tool that can deal with bot detection challenges and adapt to the constantly changing web data gathering landscape.

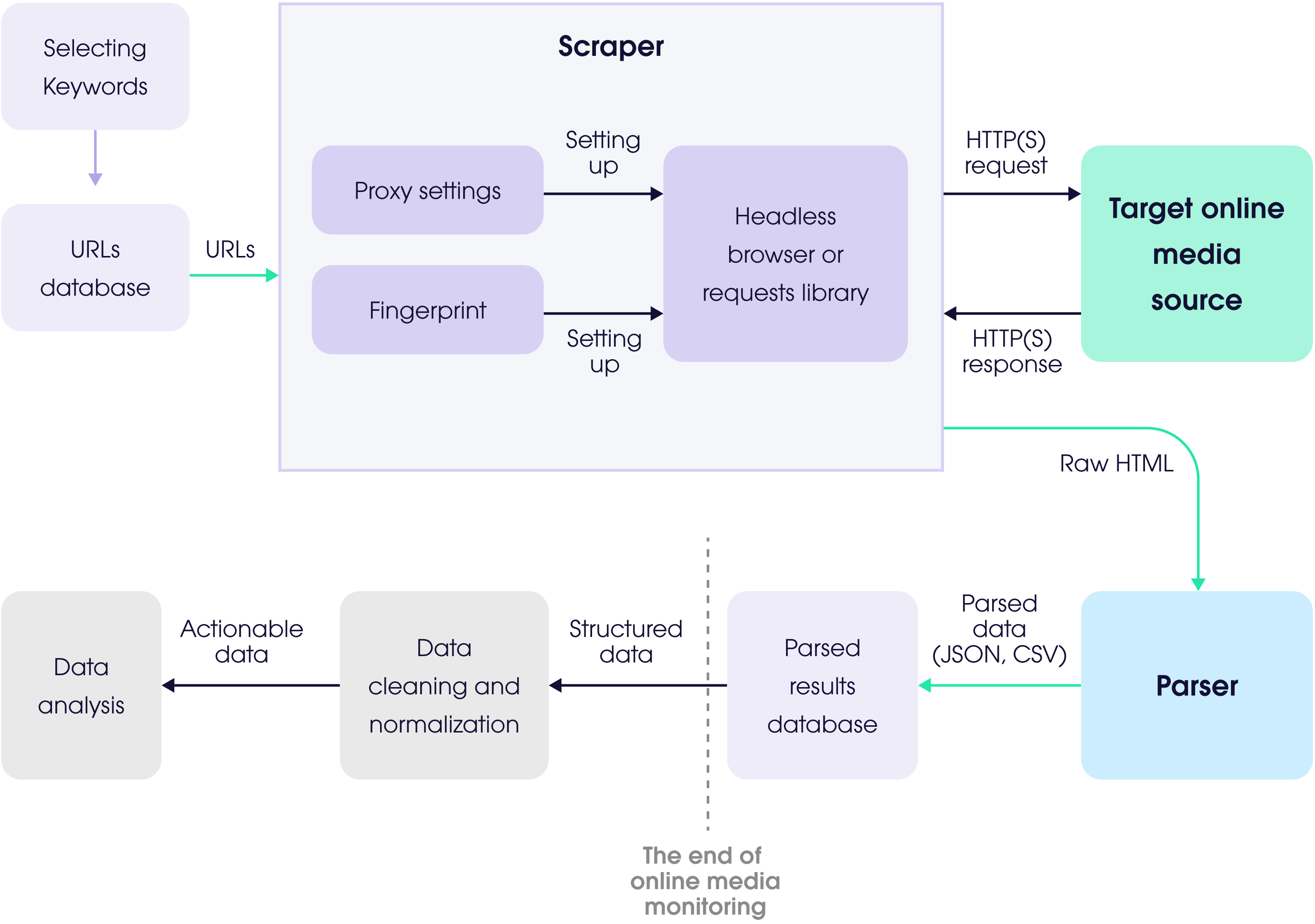
3. Collecting real-time data

Online media monitoring is time-sensitive, meaning you have to ensure updates from the required sources at a specific frequency. For example, if you want to efficiently deal with customer feedback, especially negative ones, you will need to arrange an uninterrupted data flow at specific time intervals.

Online media monitoring infrastructure

Considering the challenges, online media monitoring requires building a reliable but, at the same time, flexible infrastructure. It encompasses five major system elements, which are:

- Selecting keywords
- Discovering and collecting target URLs
- Setting up a web scraper
- Data parsing
- Data cleaning and normalization



Selecting keywords

All online media monitoring architecture involves selecting keywords you want to track online. Since the data gathering tools will be collecting all mentions based on the chosen keywords, they will determine the data you'll retrieve.

That's why it's essential to choose your keywords wisely and base them strictly on the topics you want to follow. Here's some examples of the keywords that you can use to monitor the topics of your interest.

- Your brand name
- Your service or products names
- Related phrases
- Your CEO and other key employees' names
- The same points listed above but applied to your competitors

Discovering and collecting target URLs

Once the keywords are selected, you'll need to think of ways to find places where they might be mentioned.

One of the main peculiarities of online media monitoring is that you never know where and when a new mention of your brand will pop up. It means that you can't solely rely on a database of pre-selected target URLs but should also run independent search to discover new places where the topics of interest may appear. In other words, in online media monitoring, there are two ways to look for keywords and HTMLs.

Collecting URLs

First, you need to create a list of URLs containing the data you need. These starting points can be news aggregator websites publishing articles and information about your industry. The best way to secure a complete URL database is to comb through the internet with automated web crawling tools to discover the most appropriate candidates.

Discovering URLs

If you don't have a URL database or want to constantly enrich it with new sources, you'll have to run a web and site search for specific keywords. For instance, you might run a search for the pages created over the last 24 hours to see what comes up. Since checking every single link costs you money, over time, you can shorten your discovery process by adding new URLs to your database and scraping them directly.

How often do you need to refresh target URLs or run a discovery process?

There might be a difference in frequency between refreshing the URLs from an established database and exploring the web for new sources. You will probably start out by doing the discovery part every time you want your data to be updated. Then, over time, you may notice that you get the same URLs every day and that new ones appear about once per week on average. At that point, you can decide to run your discovery process weekly but refresh the data in the URL list once a day.

Frequency also depends on your business needs. Suppose you're an agency providing media monitoring services, and your main purpose is to prepare reports for your customers containing the information on what articles they appeared in during the past 24 hours. In that case, there's no need to pull data more frequently than one time per day.

Additionally, it also depends on your target source. There may be one single website that changes a couple of times per day, but if it's an important one, you'll need to check it more frequently. For example, if you know that some particular media outlet updates no more than 3 times per day, there's no need to scan it any more frequently. If you explore the web searching for new mentions in the yet undiscovered corners of the internet, you might want to run it more frequently as you never really know how often it changes.

Setting up a web scraper

After everything is prepared for scraping, it's time to set up a web scraper itself.

> On a technical level, scraping is more or less the same for every business use case. However, the difference lies in your targets and how protective they are of their data. Typically, it's easier to scrape news websites than e-commerce websites, as they are not getting scraped as often as the latter ones. But, of course, it varies from case to case.

Aleksandras Šulženko
Product Owner of Scraper APIs

Setting up a web scraper includes three main steps:

- Configuring proxies
- Creating a fingerprint
- Sending an HTTP(S) request

Proxies for online media monitoring

There's no straightforward answer as to what proxies are better for scraping online media. The best approach is to experiment with various solutions and see what works best.

You can always start with a low-cost solution, such as shared or dedicated datacenter proxies. If it works, you can stop your search there, if not, you might opt for residential proxies.

Oxylabs proxy solutions for online media monitoring:

Shared Datacenter Proxies. Cost-effective solution with datacenter IP addresses shared by multiple users at the same time. These proxies enable geo-targeting and unlimited concurrent sessions.

Dedicated Datacenter Proxies. These proxies are a definite version of Datacenter Proxies, offering a full range of customizable options and capabilities. With them, you are given dedicated IP addresses that no one else uses, which eliminates the risk that your proxies will get blocked due to some other user's activity.

Residential Proxies. In case Datacenter Proxies are not handling your targets, you should opt for Residential Proxies. With the power of a huge pool of dynamically changing IP addresses allocated to Internet Service Providers (ISPs), you'll have a high chance of effectively extracting the needed public data.

Creating a fingerprint

If you’re aiming at a challenging target, you need to create a fingerprint allowing you to slip through a website’s protection mechanisms.

What is browser fingerprinting?

A browser fingerprint is information gathered about the software and hardware of a computing device for identification purposes. Simply put, it’s the data the web server collects about you while you’re trying to connect upon your visit. A fingerprint may contain such information as an IP address, headers, cookies, hardware, browser, and system data.

Creating a believable fingerprint resembling the one of an organic user is imperative if you want your scraping efforts to be rewarded. A widespread mistake, especially when you’re new to web scraping, is using peculiar fingerprinting indicators and patterns that will give the scraper away. Thus, the best tactic is to decrease your browser's uniqueness.

- Use common browsers and user agents;
- Cut down the number of plugins installed;
- Don’t allow discrepancies, e.g., an IP and browser time zone mismatch.

Recommendations may vary from case to case. Keep experimenting to find the right combination.

Sending an HTTP request

Finally, we're approaching the final step of setting up a web scraper, i.e., sending an HTTP request. An HTTP request is a combination of an URL, headers, and proxy settings sent to the target server with the help of a request library or a headless browser. Here's the list of the most commonly used request libraries:

- **JavaScript:** Request, Axios
- **Python:** Requests, aiohttp
- **Golang:** Colly
- **R:** rvest
- **Ruby:** HTTParty, Kimurai
- **PHP:** Goutte, Panther

As an alternative, you can entrust this job to an API (Automated Programming Interface), such as the SERP Scraper API and Web Scraper API.

As mentioned before, in online media monitoring, there are two ways to look for target keywords. The SERP Scraper API is a good fit for the exploration and indexing process as it's capable of discovering the needed data on search engine results pages.

Then, to make sure that the discovery part was successful and the results retrieved contained the required information, you would want to visit those pages and verify them. For this, you can use the Web Scraper API.

Data parsing

Typically, the data extracted straight from the web is raw and unstructured. To make sense of it, you need to use parsing technologies. In the most basic sense, a parser is a technology capable of identifying the needed data in the HTML string and converting it into a readable format, such as JSON, CSV, etc.

In a broader sense, parsing is the process of drawing meaning from the retrieved data. The parsing methods may vary depending on the rules you create yourself. Based on the result that you want to achieve, you may use different sets of rules to process the text or an HTML.

Parsing can be simple when, for example, an online media monitoring agency has to provide its clients with a report containing the URLs where their clients were mentioned. In this case, a URL is the data point you already have, even before scraping the page as you got it while exploring. You can also retrieve such data points as the publication date and time, author, H1 tag, metadata, etc.

But when we're talking about online media monitoring, there can be way more complex requirements, such as measuring the sentiment of the text and looking into the context in which the brand was mentioned. For this, more sophisticated parsing rules are needed, involving AI and ML algorithm-based technologies.

What is sentiment analysis?

Sentiment analysis is a natural language processing method used to determine sentiment in a piece of text. This automated technique helps businesses grasp and analyze feedback and opinions regarding a certain brand and its products.

These are some examples of the information from online media that can be parsed depending on your parsing requirements and methods:

- URLs
- Publication date and time
- Author
- H1 tag
- Metadata
- Sentiment scores

Data normalization (optional)

Data normalization is the final stage of building an online media monitoring infrastructure. Simply, it is the process of cleaning the collected data to optimize management.

The major goals of data normalization are to get rid of duplicates and group data in a logical and coherent manner. After the normalization phase, the data can be forwarded for subsequent analysis.

Conclusion

Online media monitoring is truly a goldmine for insights into a brand's public image, reputation, and overall development. It should be given proper credit and treated as one of the essentials that help businesses create a favorable impression on both audiences and investors.

This responsible task requires a solid foundation to rely on. Building and maintaining a real-time online media monitoring infrastructure may seem cumbersome, but at the end of the day, it will pay off.

Even though online media monitoring infrastructures may drastically vary depending on a project's requirements, there are key elements integral to any architecture of this kind. They are: selecting keywords, discovering and collecting target URLs, setting up a web scraper, data parsing and normalization.

Hopefully, knowing these stages and general guidelines on how to undergo them will help you build your own real-time online media monitoring infrastructure.

oxylabs®

Want to Know More?

If you would like to know more about any of the topics mentioned in this white paper or **learn about our products**, please get in touch! Our team is ready to answer any of your questions and offer you the best solution for your business needs.

Get in touch with Oxylabs

Our Mission

Our mission is to share all the know-how that we collected over the years in the industry in order to create the future where big data is accessible to all businesses. We seek to create a healthy environment for everyone to grow and thrive in.

Our Values

As a leading company in the proxy and web scraping industry, we ensure that the highest standards of business ethics lead all our operations. Our core values guide us toward achieving our mission. **Learn more**