



Meeting the Increasing Business Demand for Web Intelligence

Glossary

Web scraping is the process of extracting data from the web using specialized software tools called web scrapers.

Web crawling is traversing a website (or multiple websites) by using links between the pages. In web scraping, web crawlers can catalog the URLs of web pages that will be used for data extraction.

Crawling logic is a set of predefined rules and parameters that a web crawler follows when crawling the Internet.

Data parsing is the process of transforming data from one format to another, typically, in order to change its structure and prepare for analysis. The software tools used for this purpose are known as **data parsers**.

Web intelligence refers to data and insights derived from data extracted from the web and used by organizations to improve decision-making and other processes.

A web scraping pipeline is the full setup of tools and processes that create consistent automatic data flow from the Web and into the organization's data storage units.

Target unblocking is the process of circumventing geo-blocking and anti-bot measures to collect public web data from challenging targets.

Introduction

Web scraping, the automated extraction of publicly available data from the web, has become one of the most important data sources for businesses. The increasing global demand for web scraping reflects this. Growing at 15% CAGR, the web scraping software market is [expected to reach](#) a valuation of almost 1.5bn USD by 2033.

Real-time web intelligence helps companies stay competitive, optimize operations, and improve short-term and long-term decision-making. Businesses utilize various data points, from pricing and product description data to open-source information about cybersecurity threats and market sentiment data. One use case for web data that is ceaselessly gaining traction after the generative artificial intelligence (Gen AI) boom is training machine learning (ML) algorithms for various analytical and operational tasks.

Developers responsible for in-house web scraping operations face many challenges ranging from a lack of resources to the relentlessly improving anti-scraping measures. Identifying precisely what areas of building and maintaining scraping pipelines create the biggest obstacles to timely data acquisition is crucial for effectively addressing these challenges. To shed more light on web data collection intricacies, Oxylabs and Censuswide surveyed web scraping professionals who deal with these challenges every day. The results and insights of the survey are presented in this paper.

Many companies today are turning to AI to overcome various business performance challenges. According to [Forbes Advisor](#), 64% of business owners anticipate that AI will improve productivity in their company. The web scraping industry is no exception. While public web data is necessary for gen AI tool development, AI solutions, in turn, can help collect such data. Thus, the research discussed below also looks into the trends of AI usage for web scraping.

“The rapid development of AI capabilities opens new paths for technological improvements. Thus, while focusing on the challenges and business costs associated with public web data extraction, we also used this survey as an opportunity to research AI adoption by web scraping professionals. We are certain that sharing the insights drawn from the data we collected will help stakeholders advance the field of web intelligence and their own goals.”



Julius Černiauskas,
Chief Executive Officer at Oxylabs

Survey methodology

Oxylabs, in partnership with research consultants Censuswide, surveyed software engineers and other professionals whose work involves web scraping. In total, 506 professionals participated in the survey. The workplaces of the respondents are nearly equally divided between two of the biggest markets for public web data gathering, the United States of America (USA) and the United Kingdom (UK).

The survey participants work in companies of all sizes, ranging from less than 10 to more than 500 employees. The annual turnover of these businesses ranges from under £100,000 to £500 million or over, showcasing that businesses of all sizes utilize web data today.

The survey comprises 14 questions covering the following general topics:

- Did the demand for web data increase over the last 12 months?
- What challenges do scraping professionals face when collecting web data?
- Is building and maintaining data parsers a source of notable challenges and business costs?
- What is the business impact of delayed data collection, and how soon is it felt?
- Do scraping professionals utilize AI solutions to address these challenges, and are they satisfied with the results?

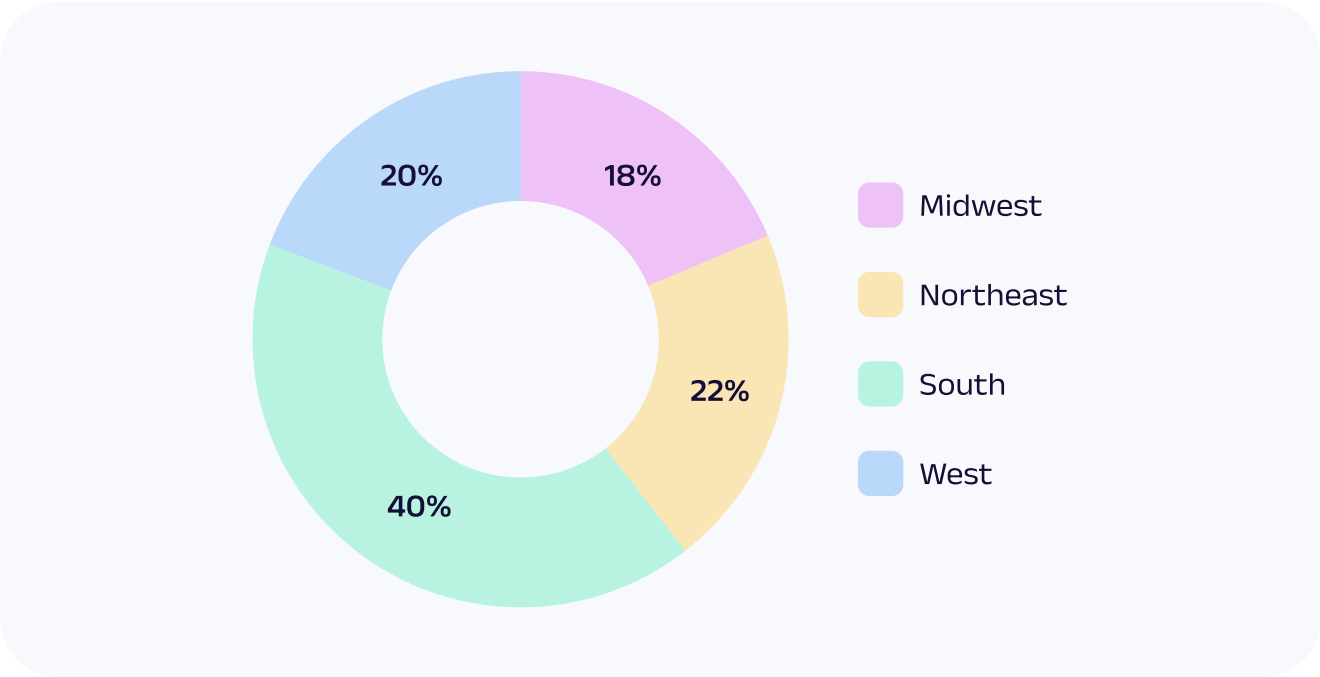
The research aimed to identify the issues that people responsible for public web data acquisition in their companies face and which of these issues have the most significant impact on overall business results. Oxylabs' developers recognized parsing as a repetitive and time-consuming task that often puts a strain on resources devoted to scraping. Thus, part of the survey puts an emphasis on building and maintaining data parsers to test the hypothesis that this area provides special opportunities for improvement.

The insights summarized here are by and for software engineers, full-stack developers, back-end developers, data engineers, data scientists, python developers, and other professionals who will be able to evaluate how their experience reflects or strays from that of their peers. The answers to the questions will also help business leaders and web scraping solutions providers plan for addressing the main challenges.

Survey sampling

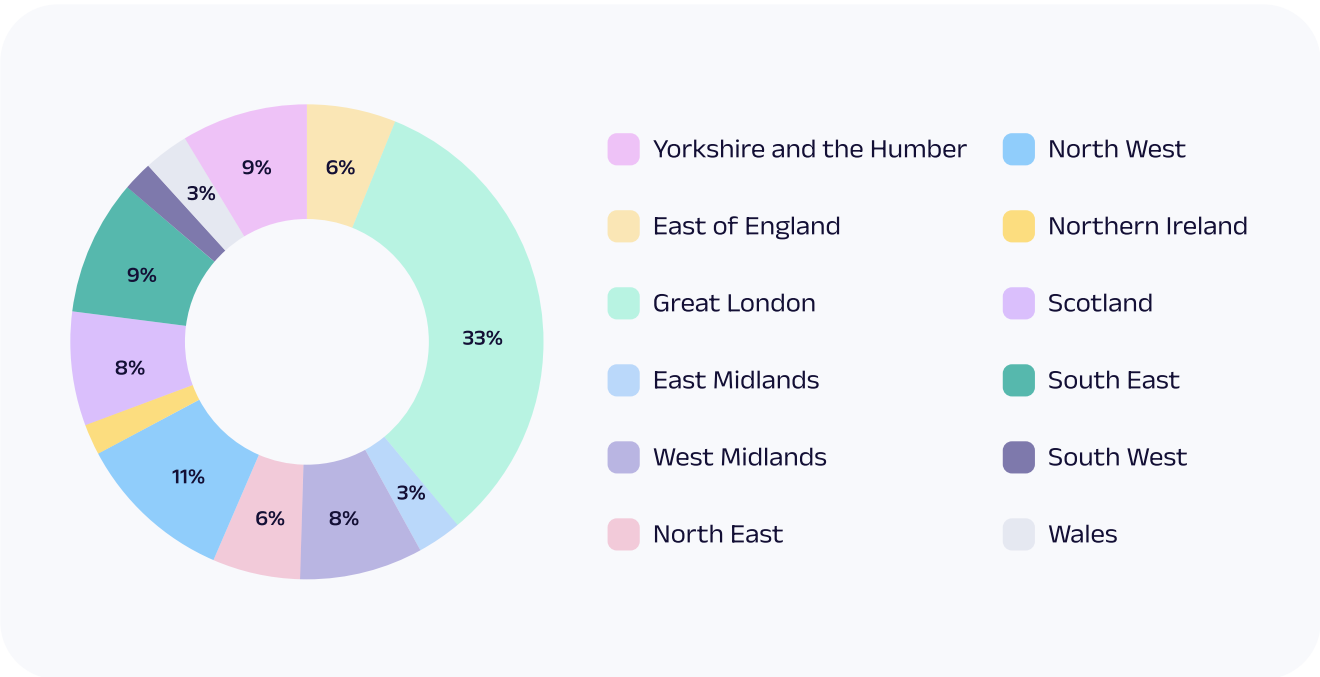
USA respondents regional spread

Where do you work?



UK respondents regional spread

Where do you work?

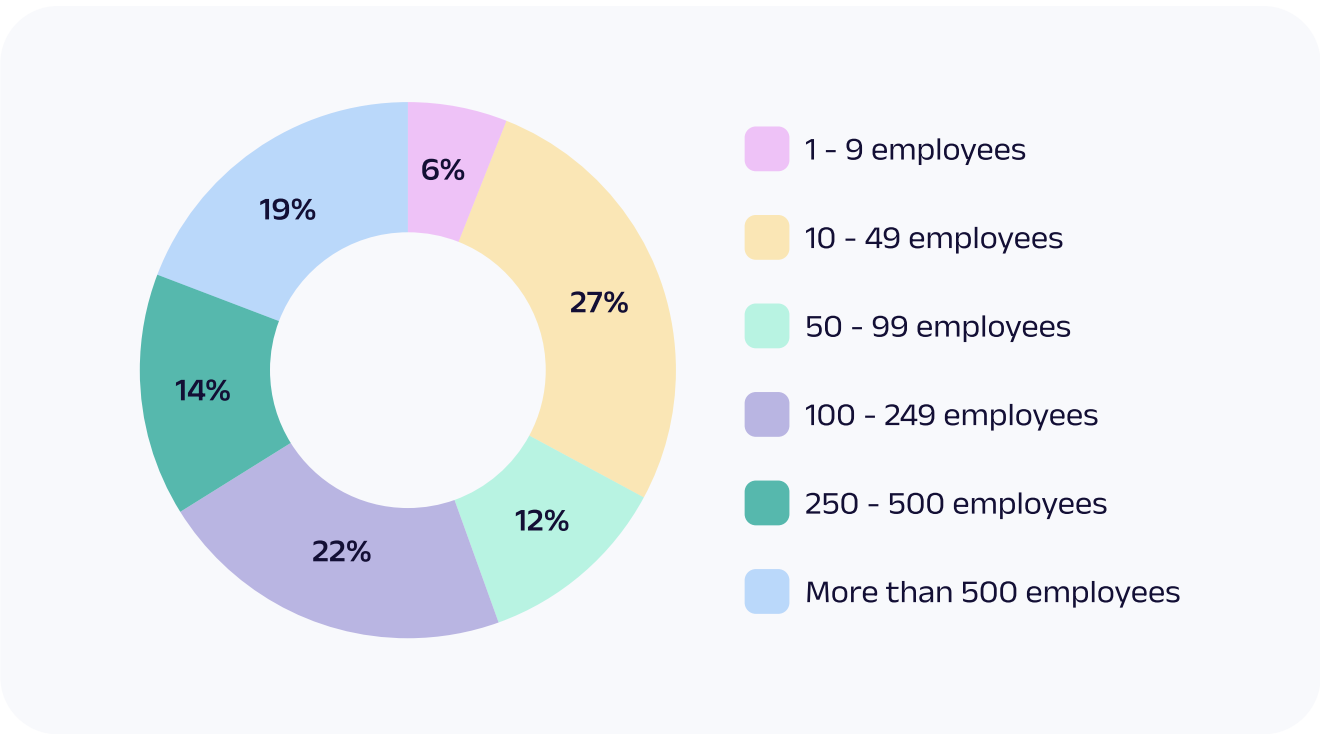


Job titles of the respondents

Which of the following best describes your job title?	Percent
Software Engineer	41
Back-end Developer	3
Full-stack Developer	22
Data Engineer	20
Data Scientist	9
Python Developer	2
Other	3
Number of respondents	506

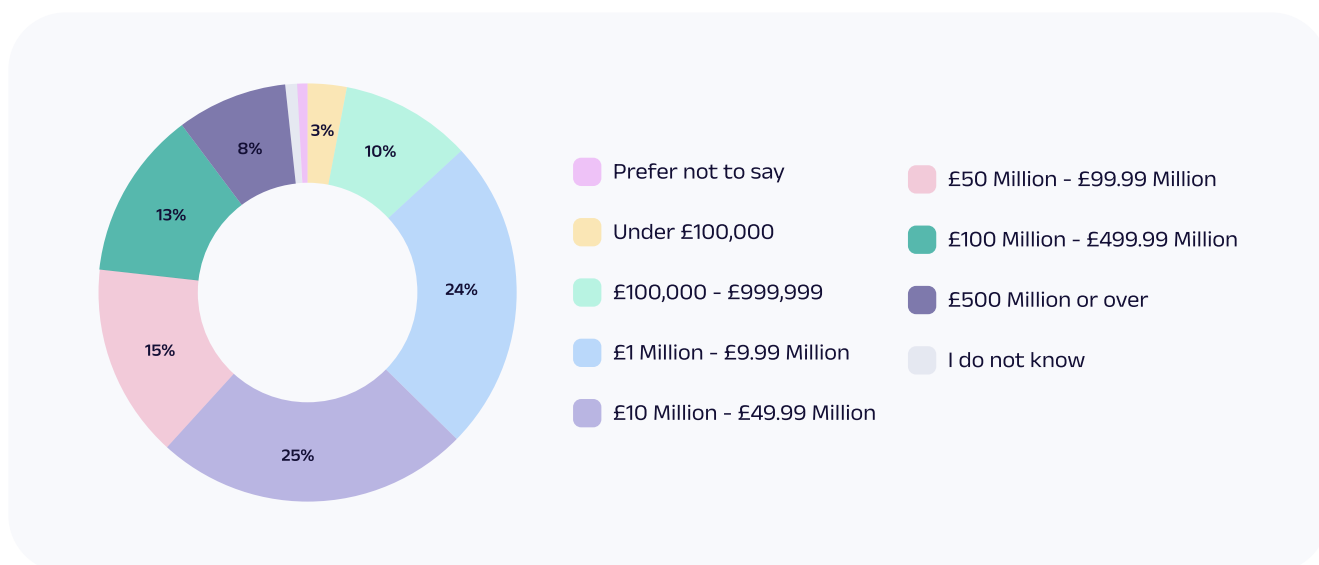
Company size

What is the size of the company you currently work for?



Company annual turnover

Approximately what is your company's average annual turnover?



Key trends and insights

- Over the last 12 months, the need for public web data increased **for 74%** of US and UK businesses. **For 24%**, the increase was **significant**.
- **61%** of respondents identified building the necessary infrastructure as the main challenge of large-scale web data collection.
- Building and maintaining data parsers was the second most recognized challenge, named by **half** of the respondents, with **86%** acknowledging that it is a time- and- resource-intensive task.
- **98%** of businesses face some **negative impact** when data collection is delayed. For 21%, the impact is **severe**.
- If parsing is interrupted, **95%** of businesses face negative impacts **within 24 hours**.
- **75%** of developers spend **10 to 40 hours** weekly on parsing processes.
- Identifying complex parsing patterns when dealing with multiple URLs is the most common challenge (**58%**) associated with data parsing, with dynamic website layouts and changing structures (**56%**) being a close second.
- **57%** say that due to changing website layouts, they have to fix parsers several times a week. For the same reason, **31%** fix parsers **every day**.
- **Over 50%** of developers mention **time** as the main parsing-related business cost.
- **77%** of respondents have tried using **AI-powered web scraping solutions**. Of them, **86%** say that it helped them save time and resources.

UK vs. US

Web scraping professionals in the UK and the US generally share their experiences researched in this survey. However, a few slight but notable differences have been observed:

- Developers working in the US are significantly more likely to use AI for web scraping regularly (**56% vs. 38%**).
- The share of professionals who have to **fix parsers every day** due to changing website structures is notably **greater in the US (38%)** than in the UK (**24%**).

“The survey confirms what Oxylabs observed over the past year – the demand for high-quality public web data is accelerating. With the advancements in AI model development and increased usage of tools based on these models, we expect this trend to continue well into the future.”



Julius Černiauskas,
Chief Executive Officer at Oxylabs

The main web scraping challenges relate to infrastructure building and maintenance

What are the challenges you face when collecting web data at scale? (Select all that apply)	Percent
Building the infrastructure necessary for web data collection (e.g., servers, computing power, etc.)	61
Building and maintaining data parsers	49
Maintaining and adjusting scraping pipelines	44
Lack of professionals with direct experience and skills in web data acquisition	44
Bypassing anti-scraping measures	38
There are no challenges	5
Other, please specify	0
Number of respondents	506

Securing enough computing power and otherwise building the infrastructure necessary for web data collection is the top challenge (61%) for scraping professionals. Some other major challenges are related to developing the necessary software, such as building and maintaining data parsers (49%), maintaining and adjusting scraping pipelines (44%), and building the crawling logic (33%). For 44% of respondents, the issues are amplified by the struggle to find professionals with direct experience and skills in web scraping.

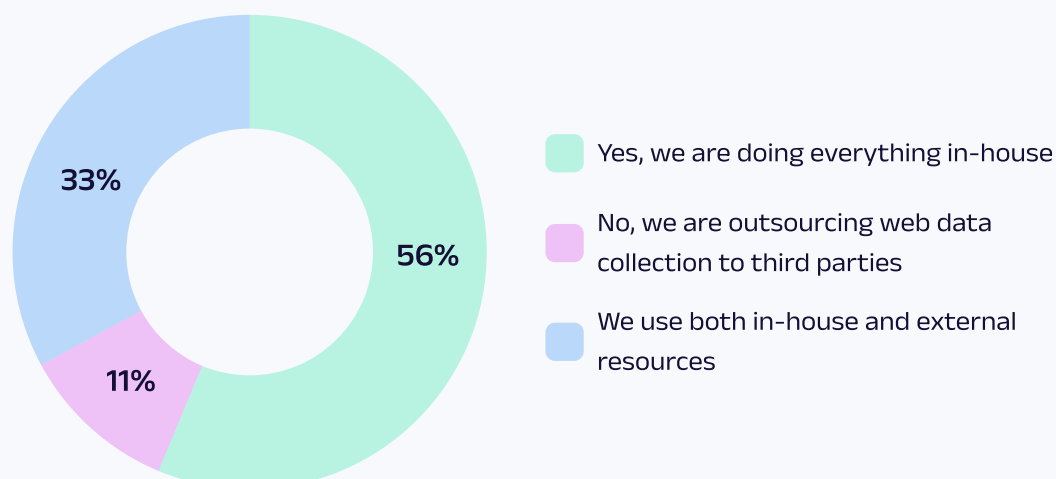
Additionally, most respondents (89%) confirm that building and maintaining web data pipelines is time- and resource-intensive, while 39% strongly agree with this claim.

To what extent do you agree or disagree with the following statements: Building and maintaining web scraping pipelines for multiple websites is a time and resource-intensive task for you.

	Percent
Agree (Net)	89
Strongly agree	39
Somewhat agree	51
Neither agree nor disagree	9
Somewhat disagree	2
Strongly disagree	0
Disagree (Net)	2
Cannot answer	0
Number of respondents	506

Despite these struggles, most of the businesses (56%) rely solely on internal resources for web data collection at scale. Only 11% outsource everything, while a third of businesses use both internal and external resources. This suggests that there is plenty of business to capture for external web scraping solutions and service providers if they can convince companies to explore third-party-based web data-gathering options.

Do you use internal or external resources for web data collection at scale?



Although the constantly improving anti-scraping measures are often discussed in the web scraping community, they have been named as an issue by a slightly smaller number of respondents (38%) than most of the other issues. Thus, while web unblocking remains crucial for large-scale solutions providers, they also need to devote more attention to helping clients streamline scraping pipelines and reduce the need for internal resources.

“Web scraping infrastructure and solutions providers are often better positioned to optimize their operations. This is mutually beneficial for the providers and their clients. The providers might attract more clients by offering better value while the clients can cut business costs by switching to better-optimized external services.”



Žydrūnas Tamašauskas,
Chief Technology Officer at Oxylabs

Data parsing puts a strain on the company’s internal resources

Building and maintaining data parsers is, at least to some extent, a time and resource-intensive task for 86% of scraping professionals.

To what extent do you agree or disagree with the following statements: Building and maintaining data parsers is a time and resource-intensive task for you.	Percent
Agree (Net)	86
Strongly agree	38
Somewhat agree	48
Neither agree nor disagree	9
Somewhat disagree	4
Strongly disagree	1
Disagree (Net)	5
Cannot answer	0
Number of respondents	506

The challenges that lead to business costs come from multiple sources as per the data below.

What are the challenges you face when building complex data parsers? (Select all that apply)	Percent
Identifying complex parsing patterns when dealing with multiple URLs	58
Dynamic website layouts and changing structures that result in broken parsing pipelines	56
Debugging the parsers and noticing errors in data	46
Extracting nested information	42
There are no challenges	5
Other	0
Number of respondents	506

Dealing with multiple URLs (58%) and dynamic website layouts and changing structures (56%) pose issues for most of the respondents. Nearly half (46%) of the respondents say that debugging the parsers and noticing errors in data is a source of issues, while 42% struggle with extracting nested information.

The biggest business costs that result from building and maintaining data parsers come from: (Select up to 3)	Percent
Maintaining the necessary infrastructure (servers, etc.) in-house	57
Devoting developers' time for parser maintenance	54
Data inaccuracies due to parsing errors	51
Devoting developers' time for parser building	48
There is no biggest challenge	5
Other	0
Number of respondents	506

When we turn to the biggest costs related to data parsing, we can see that, once again, in-house infrastructure is the main issue area. As much as 57% say this is where the lion's share of parsing-related costs comes from. Around half (51%) of the respondents name parsing errors as a costly issue. Developers' time needed to build (48%) and maintain (54%) data parsers is another major source of business costs. Interestingly, parser maintenance is recognized as cost-inducing even more than the initial building of parsers.

One explanation for this is that parsers constantly break due to changing website structures. This leads 31% of developers to fix parsers every day, while as much as 57% fix them at least several times a week.

Because of the changing website structures, you have to fix data parsers	Percent
Constantly, even every day	31
Often, up to several times per week	57
Rarely	12
Never	1
Number of respondents	506

On the whole, building and maintaining data parsers for most developers (57%) costs 11 to 20 hours per week. One-quarter of the respondents say that they dedicate up to 10 hours a week to these tasks. Meanwhile, 18% spend between 21 and 40 hours per week on parser building and maintenance.

“Most of the time, a parser built for one domain will not work on the other. Additionally, websites constantly update their layouts, after which your parser will not find data in the same place it used to, and the parsing process will break down. Thus, both the scope of the task and the fact that it is never really done constantly take away the developer’s time. And both need to be addressed in order to free up more of that time for more sophisticated tasks.”



Aleksandras Šulženko,
Product Owner at Oxylabs

Thus, there seem to be two main approaches for companies who wish to optimize their web intelligence collection costs. One is to outsource the infrastructure and scraping processes, moving the related costs to a third party if a good enough deal is available. For those who wish to continue relying on in-house infrastructure, it is crucial to find a way to save developers' time spent on parser building and maintenance. This second option is made more challenging by the fact that parsing errors already induce additional costs, and reducing the time spent on this task is likely to increase the number of errors. Therefore, the second approach relies on the possibility of implementing automated solutions that consistently produce results with a tolerable margin of error.

The importance of timely web intelligence for business is growing

Businesses might be especially alert to the urgency of solving the main web scraping issues now as getting the necessary web data on time is becoming increasingly critical. Firstly, as noted, the demand for web data has increased over the past twelve months, with 49% of respondents noting a slight increase and 24% stating that they noticed a significant increase.

Have your public web data needs increased, stayed the same, or decreased over the past 12 months?	Percent
Increased (Net)	74
Increased significantly	24
Slightly increased	49
Stayed the same	23
Slightly decreased	2
Decreased significantly	0
Decreased (Net)	3
Cannot estimate	1
Number of respondents	506

The fact that the demand for web data increased to some extent for 74% of businesses tracks a more general trend, as increasing investment in data has also [been reported](#) by most companies in 2022 and 2023. There seem to be no identifiable indications that the upward trend of increasing demand for data in general and web data, in particular, will subside soon.

What is the impact on your business if data is not collected as scheduled?	Percent
Severe—it badly affects business decisions	21
Moderate—it delays some processes	67
Minor—there is little impact	10
We observe no impact	3
Number of respondents	506

It is crucial to note that most web intelligence collected by companies is time-sensitive. If data is not sourced as planned, 67% of businesses experience a moderate negative impact, while for 21%, this impact is severe.

Most companies (77%) feel these negative effects within 1 to 10 hours after the data flow is interrupted. For 7% of businesses, these effects occur in less than an hour.

In case the data flow is interrupted by broken scraping and parsing pipelines, how soon does this start to impact your core business processes, if at all?	Percent
In less than 1 hour	7
1-5 hours	37
6-10 hours	40
11-24 hours	12
More than 24 hours, please specify %s hours	1
Cannot estimate	4
Number of respondents	506

Thus, it is not only data but the speed and reliability of its collection that is becoming more and more critical. These findings confirm what could have been hypothesized about public web data — since the web is an extremely dynamic data repository, information can age at a previously unseen rate. From this, we can predict that the importance of real-time data collection and reliable infrastructures will continue to grow.

“The dependence on real-time, uninterrupted data flow might explain why companies hesitate to outsource their web scraping tasks. You want to have control over procedures that have a quick and direct impact on your business goals. If external service providers can offer flexible services that maximize client control while taking off the load related to reliability assurance, they might be able to win these businesses over.”



Tomas Montvilas,
Chief Commercial Officer at Oxylabs

The usage of AI-powered solutions saves time and resources

Over the last few years, various industries have decisively turned to Artificial Intelligence (AI) as the main source for potential improvements. Web scraping is no exception. Almost half of scraping professionals (47%) use AI-powered solutions regularly. In total, 77% have tried using AI solutions, while another 20% would love to try them. This leaves only 3% of web data collection professionals who have no interest in utilizing AI at the moment.

Are you currently using any AI-powered web scraping solutions? (Select all that apply)	Percent
Yes, I am using them regularly	47
Yes, but do not use them regularly	26
No, but I would love to try one	20
No, I tried them but they didn't meet my expectations	4
No, I don't think I need them	3
No, other, please specify	0
Number of respondents	506

Thus, generally, it is reasonable to say that the web scraping community welcomes AI innovations and is prepared for the change these innovations can bring to daily work. And the results so far are promising. Only 4% of the respondents say that AI did not meet their expectations. Meanwhile, 86% of those who tried AI solutions for web scraping confirmed that it helped them save time and resources. Thus, the trend of AI adoption is likely to continue.

In case you tried using AI-powered web scraping solutions, have they helped you save time and resources?	Percent
Yes	86
No	8
It is hard to evaluate the impact	6
Number of respondents	390

“In many industries, the true extent of AI’s potential impact and its ability to provide a return on the huge investment it had already accumulated is still up for debate. However, large-scale web data collection seems to be perfectly suited for AI enhancements. Building and maintaining data pipelines consists of multiple operations and repetitive tasks that AI solutions can automate or improve. While we are still at an early stage of AI adoption, it is both exciting and vital for web scraping companies to closely watch the emerging AI/ML innovations and trends of leveraging these tools.”



Juras Juršėnas,
Chief Operations Officer at Oxylabs

AI is perceived as potentially most helpful for target unblocking and building parsers

In case you tried using AI-powered web scraping solutions, for which specific tasks have you employed them? (Select all that apply)	Percent
For building data parsers	67
For target unblocking (e.g. proxy management, building dynamic fingerprints)	59
For building crawling logic	52
Other, please specify	0
Number of respondents	390

Building data parsers stands out as the most popular among the specific web scraping tasks where AI was applied. Of the 390 respondents who have tried using AI-powered tools, 67% applied them in this area. Target unblocking (59%) and building crawling logic (52%) were also named by more than half. This suggests that many respondents tried AI in at least two of these three areas. Thus, it seems that AI adopters in web scraping are testing the hypothesis that AI can streamline many, if not most, tech tasks.

In your opinion, for which specific tasks an AI-powered scraping solution would be the most suitable?	Percent
For target unblocking (e.g. proxy management, building dynamic fingerprints)	43
For building data parsers	32
For building crawling logic	26
Other, please specify	0
Number of respondents	101

The responses to the question where AI-powered boosts seem most fitting switch target unblocking with building parsers. A big share of respondents (43%) see the former task as the most suitable for AI applications. Meanwhile, building parsers are recognized as such by 32%. Another 26% believe that building crawling logic is where AI can be most beneficially applied.

Thus, while AI was tried more for building parsers, scraping professionals seem to believe that it might be even better suited for target unblocking. One explanation might be that web unblocking is a much more sophisticated task and thus might seem a more fitting challenge for the newest technological developments, such as AI tools.

This might also suggest that despite generally positive feedback regarding AI usage, so far, its application in parsing has not produced revolutionary results that would add exceptional value. The fact that parsing remained time-consuming and challenging even when AI was applied seems to indicate that there was no exceptional AI-based solution for parsing in the market at the time of this survey. Oxylabs has now introduced OxyCopilot to assist with scraping and parsing, and the initial results have been very promising. This AI-powered solution and its iterations based on client feedback will fill this market gap.

“Web unblocking was one of the first areas where we at Oxylabs utilized AI technologies, and our clients are very happy with the results. However, while web unblocking becomes a challenge in specific, although important, cases, parsing is a crucial part of each and every scraping pipeline. Thus, an AI solution for building parsers is even more widely applicable and brings constant value to the customer.”



Aleksandras Šulženko,
Product Owner at Oxylabs

Our solution – AI-powered copilot for web scraping

Oxylabs has developed the AI-powered [OxyCopilot](#) as a unified [Web Scraper API](#) feature to address some of the main web scraping challenges. As our clients are now using it, we are continuously gathering feedback to determine how it can be made even more effective.

OxyCopilot is currently made up of two distinct features that can together streamline the scraping process for developers.

- Custom parser builder allows data parsers to be built in minutes. The user only needs to provide a URL of the target domain and tell the tool in natural language terms what data to return. The tool creates a parsing template for the API, which will then return the parsed data.
- Request builder helps create initial scraping requests after providing only minimal parameters. Thus, the user does not need to know all the documentation of Web Scraper API. After processing the request with a URL and natural language instructions, the AI-powered tool will generate the full scraping request with all the parameters in place.

With the combination of these two functionalities, OxyCopilot addresses some of the major challenges named by scraping professionals. It saves developers' time, avoids business costs created by heavily delayed data collection, simplifies onboarding for those with less web scraping experience, and reduces the needs and costs of internal computing resources.

"OxyCopilot grew out of an internal tool that we felt we needed as developers who constantly build and rebuild parsers. It came from an experience that we knew we shared with many peers. The survey confirms this hypothesis that developers welcome AI boost for parsing and other tasks. Now, it is up to real-life experiences to determine how well the copilot works and what iterations would be the most relevant to developers. Thus, as always, we will collaborate with our clients and the wider community regarding feedback."



Martynas Juravičius,
R&D Tech Lead at Oxylabs

"With OxyCopilot, we aim to address some of the most pressing issues that we recognized over the years and that have been confirmed with this survey. As a company that enables others to collect real-time data, Oxylabs firmly believes that the best decisions, innovations, and improvements are data-driven. Thus, we will continue monitoring the markets, testing our tools, and reaching out to professionals to understand their challenges and needs. Hopefully, the insights of this research shed some light on the opportunities in our industry that will help us all move forward."



Julius Černiauskas,
Chief Executive Officer at Oxylabs

Want to Try OxyCopilot?

Speed up data collection with the first AI-powered assistant designed to handle diverse website structures and meet your data extraction needs without manual coding.

- Start scraping using English language prompts
- Parse raw HTML into structured data
- Scale through the [Web Scraper API](#) platform

Try for free

Our Mission & Vision

We strive to create a future where all businesses have access to big data and business intelligence, aiming to create an environment where everyone can grow and thrive.

Our Values

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

[Learn more](#)