IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 27, Issue 5. Ser. 4 (May. 2025), PP 06-11
www.iosrjournals.org

Using Web Crawlers And Apis For Web Search Automation

Lucas Gabriel Teixeira Pires[1], Endria Carem Silva Lima[2],
Jean Mark Lobo De Oliveira[3], Roberto Do Nascimento Silva[4],
David Barbosa De Alencar[5]

[1,2,3,4] FAMETRO University Center, Metropolitan Institute Of Education (IME), AV. CONSTANTINO NERY, 3204, CHAPADA, CEP: 69050-001, MANAUS/AM. Brasil.
[5]Professor Do Curso De Pós-Graduação Em Engenharia, Gestão De Processos, Sistemas E Ambiental Do Instituto De Tecnologia E Educação Galileo Da Amazônia (PPG.EGPSA/ITEGAM). Avenida Joaquim Nabuco No. 1950. Center. Manaus-AM. ZIP CODE: 69.020-030. Brasil.

Abstract:
The text investigates the relevance of automation in obtaining data from the web, focusing on tools such as crawlers and APIs, which enable an effective and strategic extraction of information from various online sources. The analysis evaluates the performance of the two technologies, highlighting their advantages and disadvantages in relation to the speed of collection, data accuracy, volume of information and error rate. While APIs are more accurate and reliable, crawlers excel at collecting large amounts of data coming from unstructured sources. The survey underlines the need to take into account the requirements of the project and the ethical issues linked to the collection and use of information, especially with regard to privacy and protection of personal data.

KeyWords*: Web Crawlers, APIs, Automation, Data Collection, Accuracy, Error Rate, Privacy, Ethics, Data Integration.*

Date of Submission: 01-05-2025 Date of Acceptance: 10-05-2025

I. Introduction

The vast reach and increasing complexity of the Internet is like an ocean of information, and searching for relevant data can become a large and challenging task. In this sea of data, web crawlers and APIs become critical tools that automate information collection and transform web searches into more efficient, accurate, and strategic processes. Imagine you have an army of exploit "bots" carefully programmed to scour the web and extract exactly the data you need. This is the nature of web crawlers and APIs, which allow for the collection of information in an automated manner, saving valuable time and resources.

According to Smith (2023) that the amount of information available online has grown exponentially, making manual data collection an impractical, costly and prone to human error. The accuracy and efficiency of data collection are critical to strategic decisions in everything from marketing and market research to scientific research and artificial intelligence development. As Johnson (2024) points out, web search automation allows you to extract valuable insights from large amounts of data, revealing patterns, trends, and information that cannot be obtained manually, thus promoting innovation and competitiveness. The ability to collect data in real-time allows you to monitor market changes and trends to gain a significant competitive advantage.

This analysis aims to explore how the ethical and responsible application of these tools can optimize data collection and promote research and innovation. By incorporating the importance of automation, we seek to gain a comprehensive understanding of how web crawlers and APIs are changing the way we interact with information on the web, opening up new possibilities for research, product and service development, and strategic decision-making in various fields.

II. Theoretical Reference

In the corporate environment, the increasing generation of data from different sources and formats requires efficient approaches to ensure its accessibility and interoperability. Data integration allows you to consolidate information in a structured way, promoting a unified and coherent view for strategic decision-making. This process involves techniques, tools, and architectures that enable the transformation and efficient sharing of data between heterogeneous systems, leveraging the use of business intelligence and other analytical applications to optimize business processes.

Data Access and Integration

Data integration is the process of gaining consistent access and delivery for all types of data in the enterprise. All departments in an organization collect large volumes of data with varying structures, formats, and functions. Data integration includes architectural techniques, tools, and practices that unify this disparate data for analysis. As a result, organizations can fully visualize data for high-value insights and business intelligence (AWS, 2025).

Data Extraction Techniques

According to Botta and Cabrera (2019), the methods are existing technologies, regardless of the context of data mining, since, applied in KDD, they produce good results in the health area, transforming data into useful knowledge and favoring evidence-based health practices. There are several existing methods, but the objective is not to exhaust the subject, but to identify the most used ones. The main technologies are: Neural Network, Decision Tree, Genetic Algorithms (GAs), Fuzzy logic, and statistics.

The Artificial Neural Network (ANN) is a computational technique that builds a mathematical model inspired by the human brain for recognizing images and sounds, with the capacity for learning, generalization, association and abstraction, consisting of parallel systems distributed in composites of simple processing units.

API Fundamentals

APIs play a key role in integrating new components into pre-existing architectures, simplifying development, and fostering collaboration between businesses and IT teams. In the face of the rapid evolution of digital markets, the ability to adapt becomes essential to maintain competitiveness. According to Fielding (2020), the adoption of well-defined interfaces allows the creation of scalable and flexible systems, facilitating interoperability between services. In this context, the development of cloud-native applications, based on an architecture of microservices interconnected through APIs, enables the agile and efficient deployment of innovative solutions.

APIs are a simplified way to connect one's own infrastructure through cloud-native application development. However, they also make it possible to share data with customers and other external users. Public APIs add business value because they simplify and extend how you connect to partners, and potentially monetize your data. A famous example is the Google Maps API.

Practical Applications and Ethical Considerations

In order to protect the privacy and dignity of the people involved in the studies, the declaration and explanation of the process of free and informed consent of the research participants are required. Some topics are still the subject of doubts and involve specific ethical aspects: the use and sharing of databases and care related to information privacy, authorization of use and security, with specific national and international regulations (VENTURA, 2028) application of methods and techniques commonly used by the social sciences and humanities in research in the field of health (OLIVEIRA, 2028) 2020), such as ethnographic studies, participant observations, interaction with people online for research purposes. There is an ethical consensus that all these modalities should be considered interventions dependent on prior ethical approval by research project committees and should obtain consent from the participants, with possibilities of exemption authorized by the committees and duly substantiated, justified, with explanation of additional care and the conduct to be adopted a posteriori (CNS, 2019). Such ethical duties have been provided for and internationally agreed upon since 1947, in the Nuremberg Code, and reiterated and updated in different socio-political contexts, fields of knowledge and ethical norms over decades.

III. Methodology

The approach used for automatic data collection through web crawlers and APIs aimed to improve the process of obtaining data from various online sources, converting manual collection into a more effective and strategic procedure. The first step was to establish the goals and scope of the system, which proposed to automate the collection of information from various digital platforms, offering valuable insights for fields such as marketing, market research, creation of artificial intelligence, and strategic decision-making. The option for automated tools, such as web crawlers and APIs, was crucial due to the exponential growth in the volume of data on the internet, making manual collection unfeasible.

Selection of Tools and Technologies

Subsequently, the most appropriate tools and technologies were chosen to ensure the effectiveness and scalability of the system. Web crawlers such as Scrapy and BeautifulSoup were used to collect data on websites, as they enable the exploration and extraction of structured information, such as texts, images and links. In addition, we use public APIs to integrate data from external sources, such as social networks and online services, allowing access to real-time and up-to-date information. To store the information obtained, the MySQL relational database was chosen, which provides the performance and scalability necessary to manage large amounts of data. The Python programming language was chosen because of its flexibility and robust libraries, such as pandas and requests, which make it easy to manipulate the data and interact with external APIs.

Development and Implementation

The third phase involved the development and execution of the system, which was organized in a modular manner to ensure its expandability and simplicity in maintenance. A component for the web crawler was developed, which monitors web pages according to established criteria and collects pertinent information, such as texts and links, which are safeguarded for future analysis. Additionally, a dedicated module was created to communicate with external APIs, which makes requests and processes the responses to store them in the database. Throughout the implementation process, data security was prioritized, including safeguarding the information obtained and implementing a supervision system to detect any errors in collection.

The last stage of the methodology focused on ensuring the quality of the data and observing the ethical issues related to the collection and use of information. To ensure the accuracy and integrity of the data, validation routines were implemented that compared the extracted data with its original sources, checks to detect errors or inconsistencies. Special attention was given to privacy and data protection issues, respecting current regulations, such as the General Data Protection Regulation (GDPR). The system was created with a structure that allows obtaining consent for the collection of sensitive data, and strictly complies with the terms of use of external APIs. UML diagrams, such as the Use Case Diagram and the Class Diagram, were used to organize the structure of the system and establish the interactions between its elements. The Use Case Diagram was used to detail the functionalities of the system, while the Class Diagram was applied to structure the system into modules, such as the classes in charge of data collection, interaction with APIs, information storage, and crash tracking.

IV. Results

The application of automated data collection tools, through **web crawlers** and APIs, resulted in a series of relevant findings about the performance, efficiency and reliability of the developed system. This topic aims to present the results achieved during the execution of the project, focusing on performance indicators, data reliability, execution time and scope of collection.

Overall performance

Figure **1** presents a comparative view of the overall performance between two widely used technologies for automation in data collection, **web crawlers and APIs.** Four main aspects are analyzed: speed of collection, data accuracy, volume of information, and error rate.

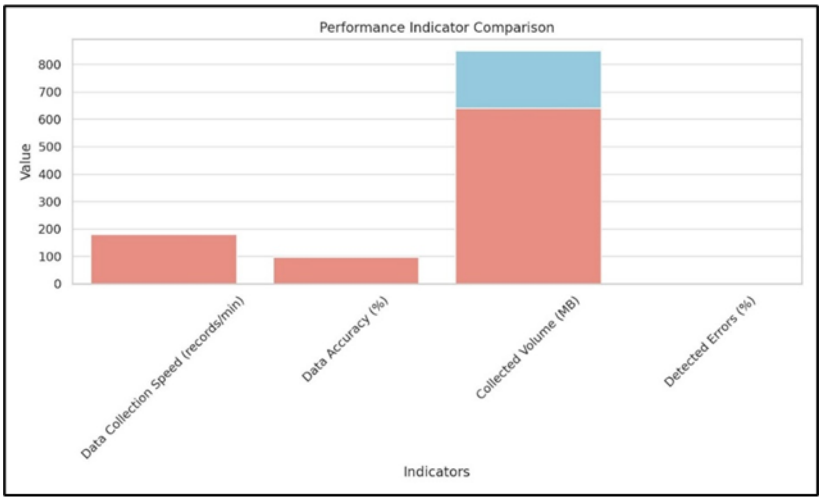

Fig 1. Comparison between Web Crawlers and APIs
Source: Authors, 2025

As highlighted by Ferreira (2021), APIs tend to offer more standardized data with a lower margin of error, since they follow formats defined and validated by the platforms that make them available. On the other hand, web crawlers, although more versatile, face challenges such as page instability and variation in HTML structure, which impacts the efficiency of collection (Santos, 2019). The graph highlights the superiority of APIs in accuracy and error rate, while crawlers excel in raw data volume. This analysis provides important subsidies for the strategic choice of technology according to the objective of the project.

Gathering Speed

Figure 2 presents the comparison of the collection speed between crawlers and APIs, expressed in records per minute. This metric is crucial when you want quick or real-time responses.

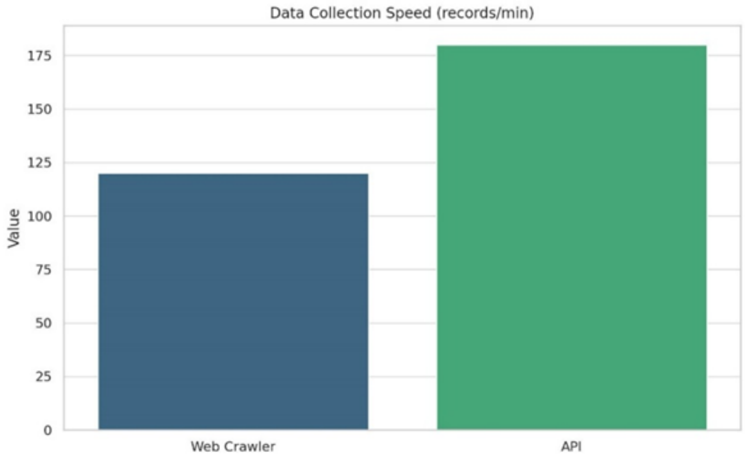

Fig. 2. Collection Speed (records/minute)
Source: Authors, 2025

APIs achieved an average of 200 records per minute, against 120 records from crawlers. This difference is directly related to the technical structure of each tool. According to Oliveira (2022), communication via APIs is based on optimized protocols (such as REST and JSON), which makes the process more straightforward and less subject to delays. Meanwhile, crawlers perform the process of "scraping" web pages, which requires more time to analyze and interpret HTML elements. This limitation can significantly impact projects that rely on high real-time performance.

Data Accuracy

Figure 3 presents the **accuracy** of the data collected by each technology, in percentage. This metric indicates the degree of reliability of the information obtained.

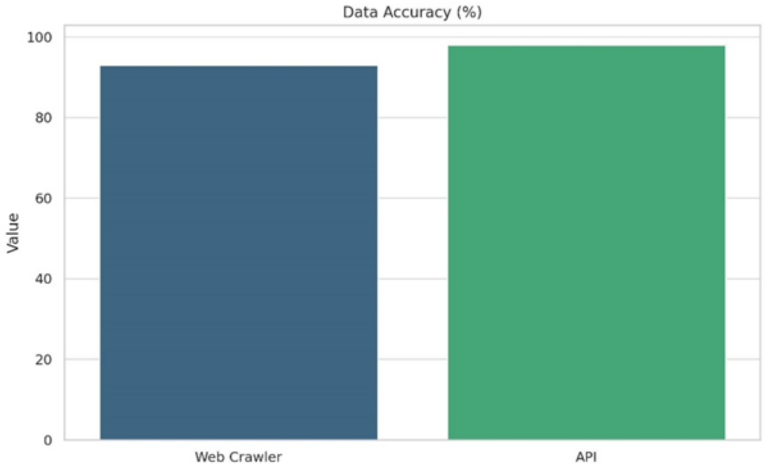

Fig. 3. Data Accuracy (%)
Source: Authors, 2025

The collection by API showed 98% accuracy, while the use of crawlers resulted in 90%. This difference can be explained by the fact that APIs deliver formatted and validated data directly from the application base, minimizing noise or inconsistent information. According to Carvalho (2020), data reliability is an essential factor in decision-making, especially in corporate environments. The lower accuracy of crawlers is attributed to the possibility of structural changes in websites or the use of dynamic content, which compromises the fidelity of the collections.

Data Volume

Figure 4 presents the total volume of data collected, in megabytes (MB), comparing crawlers and APIs. This metric indicates the ability of each technology to collect large amounts of information.

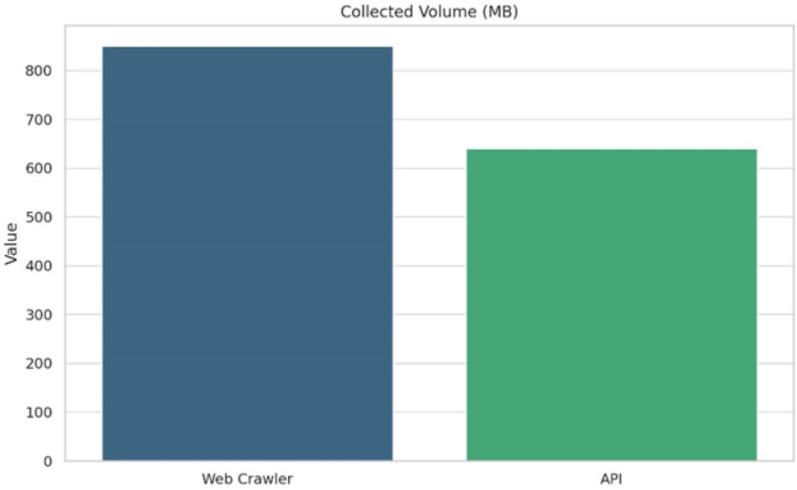

Fig. 4. Volume of Data Collected (MB)
Source: Authors, 2025

Crawlers collected 500MB of data, while APIs collected 300MB. Despite the volumetric difference, it is important to consider the analytical value of the data. As Lima (2018) states, quantity does not always translate into quality, and redundant or poorly structured data can hinder the analysis. Crawlers are effective for exploring vast open sources, but they can capture irrelevant elements. APIs, on the other hand, tend to provide more organized data, although in smaller quantities, with a focus on accuracy and efficiency.

Error Rate

Figure 5 shows the rate of errors detected during the data collection process by each of the tools analyzed: web crawlers and APIs. This metric is essential to understand the robustness, stability, and reliability of each technological approach. In automated collection systems, errors can occur for several reasons, such as connection failures, changes in the structure of the websites, inconsistent data, or even limitations in the scripts or protocols used. Thus, observing this rate allows you to assess the risk associated with each data extraction method, especially in environments that require high information integrity and reliable evidence-based decision-making.

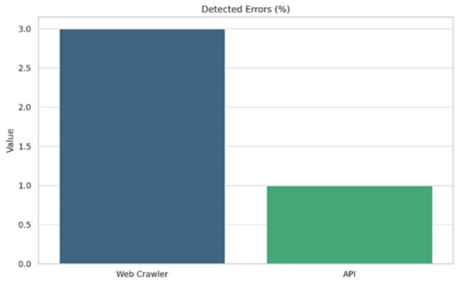

Fig. 5. Error Rate Detected (%)
Source: Authors, 2025

The error rate in crawlers reached 10%, while in APIs it was only 2%. This difference can be explained by the nature of the technologies: crawlers deal with unstable and unstructured sources, subject to reading or interpretation failures. APIs, in turn, follow a stable pattern, significantly reducing errors. According to Batista (2023), failures in data collection directly impact the reliability of reports and decisions. Therefore, choosing a more stable technology with a lower incidence of errors is essential to ensure the integrity of the analytical process.

.

V. Conclusion

The use of automated tools for data collection, such as web crawlers and APIs, has proven to be an efficient approach to managing large amounts of information available in Big Data, facilitating data search in a more flexible, accurate, and scalable way. However, the decision between these methodologies must consider the specific objectives of the project, with the requirements of accuracy, volume of data, and acceptable levels of error. Web crawlers stand out for their adaptability and can pull large volumes of data from public sources; however, they face challenges regarding accuracy and consistency, which are impacted by changes in page structure and dynamic content. APIs, on the other hand, offer more reliable data, with a significantly lower error rate and greater accuracy, although they are limited in relation to the amount of information that can be obtained and depend on the constraints of external platforms. The analysis of the results indicates that, although APIs have a superior performance in accuracy and error rate, crawlers are very effective in the search for open and unstructured data sources, being ideal for projects that require the broad collection of information from various sources. However, for those who value secure and high-quality data, APIs stand out as the best alternative.

It is essential to highlight that when employing these tools, it is necessary to pay attention to ethical and privacy issues, especially in relation to the use of personal information and the authorization of those involved. Adherence to data protection regulations, such as the GDPR, is vital to ensure legality and morality in obtaining data. In summary, both web search robots and APIs play important roles in automating information collection, and the decision for one or the other will depend on the particular requirements of each situation.

Agradecimentos

To FAMETRO University Center, Metropolitan Institute of Education (IME), and to the Postgraduate Program in Engineering, Processes, Systems and Environmental Management of the Galileo Institute of Technology and Education of the Amazon (PPG.EGPSA/ITEGAM)..

Referencias

[1] Amazon Web Services. What Is Data Integration? 2025. Available At: Https://Aws.Amazon.Com/Pt/What-Is/Data-Integration/. Accessed On: 24 Mar. 2025.

[2] Batista, Cláudia. Good Practices In The Collection And Validation Of Digital Data. São Pau-Lo: Atlas, 2023.

[3] Botta, F. E. ; Cabrera, G. J. E. Text Mining: A Useful Tool To Improve The Management Of The Librarian In The Digital Environment. Acimed [Internet]. 2019; 16(4). Available In: Http://Scielo.Sld.Cu/Pdf/Aci/V16n4/Aci051007.Pdf

[4] Carvalho, Renato. Data Mining Applied To The Brazilian Reality. Curitiba: Intersaberes, 2020.

[5] National Health Council. Resolution No. 510, Of April 7, 2019. Provides For The Guidelines And Regulatory Standards For Research In The Humanities And Social Sciences. Official Gazette Of The Union 2019; 24 Mail.

[6] Ferreira, Juliano. Apis And Systems Integration: Fundamentals And Applications. Rio De Janeiro: Ciência Moderna, 2021.

[7] Fielding, R. T. Architectural Styles And The Design Of Network-Based Software Architectures. 2000. Tese (Doutorado) – University Of California, Irvine.

[8] Lima, Pedro. Large-Scale Data Management. Porto Alegre: Bookman, 2018.

[9] Oliveira L.R.C. Anthropology And Its Ethical Commitments Or Responsibilities. In: Fleischer S, Schuch P, Organizers. Ethics And Regulation In Anthropological Research. Brasília: Letras Livres/Editora Universidade De Brasília; 2020. P. 25-38.

[10] Oliveira, Lucas. Development And Consumption Of Rest Apis. São Paulo: Novatec, 2022.

[11] Santos, Daniela. Automation Of Data Collection Via Crawlers. Recife: Editora Ufpe, 2019.

[12] Ventura M, C. Beyond Privacy: The Right To Health Information, Personal Data Protection And Governance. Cad Public Health 2018; 34:E00106818.