MUSA
Mitigating Unauthorized Scraping Alliance

Industry Practices
to
Mitigate Unauthorized Data Scraping

Mitigating Unauthorized Scraping Alliance

https://antiscrapingalliance.org

March 30, 2023

Table of Contents

I. Purpose of This Document

In this document the Mitigating Unauthorized Scraping Alliance (MUSA)[1] outlines non-binding and voluntary industry practices that promote means of detecting, preventing, mitigating, and enforcing against unauthorized data scraping on industry platforms. This document describes various types of actions by which companies may be able to mitigate unauthorized scraping.

II. Scope

Practices to protect against unauthorized scraping have significantly evolved over recent years, although there are no existing standards. This document is intended to offer suggested institutional, prevention, detection, mitigation, and enforcement measures against unauthorized data scraping.

This document also includes an Appendix with additional technical and organizational practices that companies may wish to consider adopting as part of their unauthorized data scraping mitigation strategies.

The ability and suitability for companies to implement the practices set out in this document and Appendix will depend on a range of considerations including the type of business, types of data, expectations of users, and their size and resource capacities. The inclusion of measures or practices in the document and Appendix should not be understood to mean that such practices were previously or are currently state of the art or considered mandatory within the industry.

III. Background

Unauthorized data scraping involves the large-scale collection of data available on websites and applications without the authorization of the platform or in violation of Terms of Service, typically for personal or financial gain. Scraping is distinct from hacking or a breach of security or vulnerability and cannot be equated with a data breach.

Developing practices for combating unauthorized data scraping will allow for industry to align on the best means of protecting against unauthorized data scraping.

By collaborating with regulators and industry members to build and share non-binding and voluntary practices to combat unauthorized data scraping, MUSA offers a unified front for the protection of user data. The suggested practices described in this document can help mitigate against unauthorized data scraping and describe important processes that can be maintained and updated effectively over time to serve the needs of companies.

---

[1] MUSA aims to bring together leading companies to collaborate on protecting user data from data misuse. MUSA is generating a global public dialogue on unauthorized data scraping in order to increase enforceable action against unauthorized data scraping, bring awareness to regulators and policymakers, establish industry practices and standards, promote positive and informed media, and advocate for legislative action that protects user data.

The practices outlined in this document are intended to mitigate unauthorized data scraping. However, it is necessary to acknowledge that due to the continuously evolving nature of scraping technologies and functional need for public-facing data, even comprehensive detection, mitigation, prevention, and enforcement practices can only reduce the incidence of unauthorized data scraping; they cannot prevent it altogether.

IV. Methodology

These practices have been compiled through extensive conversations with industry members and experts on measures to mitigate the risk of unauthorized data scraping. They also draw from industry research conducted by the research firm NewtonX in its study of 1300 professionals to better understand data extraction prevention.[2] The practices in this document are found in academic research, used by industry members, and represented in NewtonX's report. The practices listed below do not claim to be a fully comprehensive list of every unauthorized data scraping mitigation practice that companies may take, or to identify which measures will be appropriate for any given platform, but they offer initial guidance for potential mitigation.

V. Defining Unauthorized Data Scraping

Companies vary with respect to what types of automated data collection they allow versus what types they prohibit. Many companies face data scraping challenges on a daily basis and have their own working policies as to when scraping is or is not authorized. In some cases, companies also distinguish between data scraping, web scraping, and screen scraping. External to industry, unauthorized data scraping is often incorrectly perceived as a security vulnerability – as a theft of private data that was not meant to be disclosed. This perception is not accurate. Unauthorized data scrapers utilize functionality required for ordinary use of the platform they seek to scrape, extracting data at scale through automation. To create a baseline understanding amongst industry, policymakers, media, and the public from which to create non-binding and voluntary unauthorized data scraping mitigation practices, MUSA has attempted to generally define unauthorized data scraping as follows:

> **Unauthorized data scraping**, as referred to in this document, **is the automated collection of data at scale that violates a platform's Terms of Service.**[3]

Additional definitions of terms used in this document can be found in the Glossary of Terms in the Appendix.

---

[2]NewtonX. *[Whitepaper] Data Extraction Prevention: Best practices to combat data scraping*, https://www.newtonx.com/article/data-extraction-prevention-whitepaper.

[3]Unauthorized data scraping can also be done manually though this is much less common and beyond the scope for the purposes of these recommended practices.

VI. Unauthorized Data Scraping Prevention and Mitigation Practices

The suggested practices outlined below are intended to mitigate unauthorized data scraping. However, it is necessary to acknowledge that due to the continuously evolving nature of scraping technologies and functional need for public-facing data, even comprehensive detection, mitigation, prevention, and enforcement practices can only reduce the incidence of unauthorized data scraping; they cannot prevent it altogether.

Given the changes in the scraping landscape over time, this set of practices reflects what is currently understood to be potentially appropriate, and does not reflect historical practices.

While this document includes current suggested practices which organizations may wish to consider and work towards as appropriate, these measures will necessarily have to evolve as scraping methodologies advance.

A. Institutional

This section includes practices for establishing internal structures and support to identify and fill gaps in unauthorized data scraping mitigation knowledge in order to help platforms create an internal structure that is suited to their needs and proactively prepared to combat unauthorized data scraping threats.

1. Establish an internal knowledge system

Fill internal organizational gaps in knowledge and awareness that provide the critical information needed to allow product, engineering, policy, and enforcement teams to implement unauthorized data scraping mitigation practices.

1.1. Gain internal executive sponsorship & awareness: Ensure senior leadership is aware of relevant risks of unauthorized data scraping and of the company's unauthorized data scraping mitigation strategy.[4]

1.2. Establish main stakeholders and responsibilities: Establish a team of stakeholders from relevant departments and determine accountability and responsibilities for mitigating unauthorized data scraping.[5]

1.3. Build organization-wide awareness: Offer materials that educate employees on the organizational and user risk of unauthorized data scraping in order to establish baseline knowledge and awareness of unauthorized data scraping. Establish a process within the company that enables reporting of data scraping behaviors to relevant stakeholders.

[4]NewtonX.
[5]Ibid.

2. Understand organizational risk

These practices aim to establish practices companies can implement to understand their organization's exposure to unauthorized data scraping.

2.1. Understand scraping vector risks: Understand scraping vector risks in products. Grade the data's sensitivity, the impact if it were to be scraped, and its importance for business operations to be able to more quickly respond to unauthorized data scraping incidents. More sensitive data may require higher levels of scraping mitigation efforts.[6]

2.2. Perform risk assessments: Develop a process to assess and prioritize unauthorized data scraping risks based on principles that allow a company to weigh costs and benefits, determine allocation of resources, and target solutions based on risk and priority level.

2.3. Incorporate unauthorized data scraping risk assessment into product development: Add unauthorized data scraping risk assessment to the product development and approval process. During product ideation, design, and development, products are generally reviewed for legal, privacy, and security considerations. Adding unauthorized data scraping risks into the product development lifecycle allows for mitigation at the outset.

3. Establish mitigation policies and procedures

These practices aim to establish policies and procedures for regularly considering the risk of unauthorized data scraping in relevant organizational policies and procedures.

3.1. Update Policies: Review terms of service to ensure unauthorized data scraping and data misuse is explicitly covered as prohibited activity.

3.2. Share mitigation policies and procedures: Ensure anti-scraping mitigation procedures, policies, and decision-making are shared internally with the appropriate stakeholders from the outset.[7]

3.3. Identify or establish review procedures: Identify or establish procedures and infrastructure for reviewing possible unauthorized data scraping incidents.[8]

[6]Ibid.
[7]Ibid.
[8]Ibid.

4. Collaborate with external organizations and partners

These practices aim to increase external stakeholder awareness on unauthorized data scraping and increase collaboration between companies, while respecting antitrust and confidentiality requirements.

4.1. Share and seek knowledge: Share practices with industry members or experts to learn how to improve unauthorized data scraping mitigation techniques and promote the efficient development of industry techniques to combat unauthorized data scraping.[9]

4.2. Collaborate with industry: Join conversations about unauthorized data scraping mitigation or coalitions like the Mitigating Unauthorized Scraping Alliance (MUSA) for sharing practices and cross-industry collaboration.

B. Prevention

This section includes practices for proactive prevention and risk reduction of unauthorized data scraping. Companies should consider building a prevention strategy that balances user experience with data protection based on organizational priorities.

1. Disincentivize unauthorized data scraping

These practices aim to establish preventative technical solutions in order to create additional barriers that disincentivize unauthorized data scraping.

1.1. Remove or limit public access for sensitive products: Remove or limit public access to view content based on its sensitivity if scraped at scale. Based on content risk, require account creation, email or phone number verification, completion of anti-bot challenges to view a page (e.g., CAPTCHA, biometrics), or keep sensitive information behind a paywall to create barriers against commonly used scraping tools.

1.2. Rate limit endpoints: Rate limits restrict the amount of data threat actors can scrape from an endpoint. Determine what constitutes “normal use” and implement rate limits to restrict the number of calls a user can make to an endpoint over a specified time period. SMBs may rely on SaaS or cloud service provider systems while larger organizations may have more bespoke solutions.

[9] Ibid.

2. Predict and disrupt unauthorized data scraping events

These practices aim to establish technical solutions in order to predict and disrupt unauthorized data scraping.

2.1. Implement checkpoints: Where users or sessions show signs of bot activity, require users to re-authenticate or use tools such as CAPTCHAs to reduce bot access that may lead to unauthorized data scraping.

2.2. Prevent known threat actors: Block or limit access for known threat entities by leveraging reputation systems or other sources on IP addresses, internet service providers, cloud hosting services, or VPN services to create barriers for unauthorized data scraping.

3. Monitor and reevaluate dated or risky products and features

These practices aim to ensure that organizational code is written, reviewed, and periodically reevaluated to minimize unauthorized data scraping.

3.1. Product re-evaluation: Review and re-evaluate code, products, and features whose increased data scraping risks might outweigh user and business value. Consider updating, changing, or deprecating products or features whose current risks outweigh the benefits.

C. Detection & Mitigation

This section includes practices for identifying active unauthorized data scraping in order to be able to respond to suspected data scraping incidents or attempts.

1. Monitor and identify unauthorized data scraping

These practices aim to establish technical detection methods for unauthorized data scraping to intervene in unauthorized data scraping incidents as early as possible.

1.1. Monitor traffic: Develop, or outsource to a vendor, a monitoring system that collects data on traffic across the product architecture to improve a platform's ability to detect unauthorized data scraping.[10]

1.2. Identify indicators of unauthorized data scraping: Consider indicators that may help identify if unauthorized data scraping has occurred, such as how quickly typical users fill out forms, where on a button users click, HTTP headers and their

[10] Ibid.

order, and use features other than just IP addresses (e.g., screen size, resolution, time zone, installed fonts) to identify clusters of coordinated activity.

2. Investigate active unauthorized data scraping

These practices aim to establish an investigation process for unauthorized data scraping incidents in order to increase the speed and effectiveness of the response process.

2.1. Create a clear investigation process: Create and implement an investigation process to assess for active unauthorized data scraping. Establish a process for what occurs after suspected unauthorized data scraping is detected or when additional help may be needed to mitigate an active scraping attack which has evaded baseline defenses.

3. Remediate through technical actions

These practices aim to establish technical measures to enforce against unauthorized data scraping actors.

3.1. Revoke access: Use block lists or CAPTCHAs and make error messages nondescript where applicable to stop the unauthorized data scraping actor from further activity.

3.2. Mediate at the account level: Where feasible, consider educational rehab warnings (e.g., “It seems like you may be using automation software”), temporary lock outs, or suspending user accounts to prevent further unauthorized data scraping. This should be coupled with appropriate user appeals mechanisms.

D. Enforcement

This section includes practices for enforcement against detected and/or attributed unauthorized data scraping activity and unauthorized data scraping actors.

1. Develop an enforcement framework

These practices aim to establish a structure for identifying appropriate responses and guiding enforcement against unauthorized data scraping activity.

1.1. Response plan: Develop an unauthorized scraping response procedure that includes clear internal ownership of enforcement and response actions should internal stakeholders determine a response necessary. Not all instances of scraping may warrant a response. The decision to do so should be based on

careful consideration of relevant factors, including legal assessment where necessary.

2. Disclose identified unauthorized data scraping actors

These practices aim to establish avenues for information sharing in order to ensure that repeat unauthorized data scraping actors can be pursued through legal means where appropriate, establish transparency with regulators, and protect user data to the highest extent possible.

2.1. Share information about unauthorized data scraping actors: Consider disclosing actionable information regarding unauthorized data scraping actors to regulators and law enforcement for collaboration in responding to unauthorized data scraping incidents where appropriate.

VII. Appendix

A. Additional Practices

The included additional practices below may be effective ways to mitigate unauthorized data scraping. The ability and suitability for companies to implement the following practices will vary depending on company size, resources, and the type of user data handled by the company. There is not an expectation that companies implement any of the outlined additional practices. The following practices can be used as additional opportunities for companies to build out organizational and technical processes and strategies to combat unauthorized data scraping.

1. Fill gaps in expertise: With the team of internal stakeholders, determine gaps in expertise and knowledge and fill them with external services (e.g., external cyber security services, investigation services, operational security consultants).[11]

2. External education: Publish content related to unauthorized data scraping prevention and enforcement to educate external actors on unauthorized data scraping risks.

3. Present text content as images: Render text into an image where possible given accessibility considerations to reduce simple unauthorized data scraping and text extraction.

4. Make data aggregation more complex: Employ practices like encrypting user identifiers to make data aggregation more difficult and reduce the value of data available. Making data aggregation more complex adds barriers to unauthorized data scraping by making critical or sensitive data less easily accessible.

5. Formally request unauthorized data scraping activity to stop: Send a request to discontinue and/or a cease and desist letter to the unauthorized scraping actor, depending on the actor, jurisdiction, and severity of the unauthorized data scraping incident.

6. Pursue litigation: If an actor does not comply with requests, consider pursuing litigation and filing a lawsuit to compel compliance.

7. Establish a bug bounty program: Use an unauthorized data scraping bug bounty program to discover vulnerabilities that may not have been discovered internally.

8. Build means of detection into software: Use honeypot data and marker injections for unauthorized data scraping identification.

[11] NewtonX. *[Whitepaper] Data Extraction Prevention: Best practices to combat data scraping* (2022).

9. Review code: Automate review of code to assess vulnerabilities. Establish a process for manual code review for data scraping vectors.

10. Assess API request: Use machine learning to assess API requests for possible unauthorized data scraping activity to improve response times and notify about at-risk endpoints more quickly.

B. Glossary of Terms

API: An Application Programming Interface (API) is a type of software interface that allows two or more computer programs to communicate; a system access point or library function that has a well-defined syntax and is accessible from application programs or user code to provide well-defined functionality.[12]

API Endpoint: An API endpoint is a digital location where an API receives requests about a specific resource on its server. In APIs, an endpoint is typically a uniform resource locator (URL) that provides the location of a resource on the server.[13]

Bug Bounty Program: An initiative offered by websites, organizations and software developers that incentivizes individuals for reporting bugs, especially those pertaining to security exploits and vulnerabilities.[14]

CAPTCHA: a challenge-response test used to help determine whether a user is human or bot. CAPTCHA's are often used to help reduce spam and automated extraction of data from websites by requiring visitors to the site to solve a simple puzzle in order to gain access.[15]

Cease and Desist Letter: A cautionary letter sent to an individual or entity describing the alleged misconduct and demanding the recipient stop the alleged misconduct or illegal activity. The letter provides notice that legal action may be taken if the recipient does not cease the alleged misconduct.[16]

Honeypot: A computer security mechanism set to detect, deflect, or, in some manner, counteract attempts at unauthorized use of information systems. Generally, a honeypot consists of data that appears to be a legitimate part of the site which contains information or resources of value to attackers but is actually isolated, monitored, and capable of blocking or analyzing the attackers.[17]

---

[12] CSRC Glossary, National Institute of Standards and Technology.
[13] What is an API Endpoint?, HubSpot.
[14] NTIA, *Vulnerability Disclosure Insights Report* (2016), 9.
[15] How CAPTCHAs Work, Cloudflare.
[16] Cease and Desist Letter, Cornell Legal Information Institute.
[17] Eric Cole and Stephen Northcutt, A Security Manager's Guide to Honeypots, SANS Technology Institute.

Endpoint: Endpoints are physical devices that connect to and exchange information with a computer network. Some examples of endpoints are mobile devices, desktop computers, virtual machines, embedded devices, and servers.[18]

Rate Limiting: Rate limiting is a strategy for limiting network traffic. It puts a cap on how often someone can repeat an action within a certain timeframe (e.g., trying to log in to an account), before they are temporarily blocked from being able to repeat this action, and/or are required to complete an additional authenticity step.[19]

Scraping Vector: A scraping vector is a tactic or method that can be used to scrape data. This is distinct from an attack vector, which is a point of entry into a system that attackers may exploit vulnerabilities. Scraping vectors are distinct from hacking or a breach of security or vulnerability.[20]

C. References

Cloudflare. “How CAPTCHAs Work.” Accessed December 29, 2022. https://www.cloudflare.com/learning/bots/how-captchas-work.

Cloudflare. “What is rate limiting?” Accessed December 29, 2022. https://www.cloudflare.com/learning/bots/what-is-rate-limiting.

Cole, Eric, and Northcutt, Stephen. “Honeypots: A Security Manager's Guide to Honeypots Version 1.1.” SANS Technology Institute. March 5, 2007. https://www.sans.edu/research/security-laboratory/article/honeypots-guide.

Cornell Legal Information Institute. “Cease and Desist Letter.” Last modified November 2021. https://www.law.cornell.edu/wex/cease_and_desist_letter.

Cyphere. “What is an attack vector?” Accessed February 24, 2023. https://thecyphere.com/blog/attack-vector.

HubSpot. “What is an API Endpoint?” July 27, 2022. https://blog.hubspot.com/website/api-endpoint .

Microsoft Security. “What is an endpoint?” Accessed December 29, 2022. https://www.microsoft.com/en-us/security/business/security-101/what-is-an-endpoint.

NewtonX. *[Whitepaper] Data Extraction Prevention: Best practices to combat data scraping*. 2022.

[18] What is an endpoint?, Microsoft Security.
[19] What is rate limiting?, Cloudflare.
[20] What is an attack vector?, Cyphere.

https://www.newtonx.com/article/data-extraction-prevention-whitepaper.

National Institute of Standards and Technology. “CSRC Glossary: Application programming interface (API).” Accessed December 29, 2022. https://csrc.nist.gov/glossary/term/application_programming_interface.

National Telecommunications and Information Administration. *Vulnerability Disclosure Attitudes and Actions*. 2016. https://www.ntia.doc.gov/files/ntia/publications/2016_ntia_a_a_vulnerability_disclosure_insights_report.pdf.

///