PUBLICATION

Web Scraping and the Rise of Data Access Agreements: Best Practices to Regain Control of Your Data

Authors: Alexander Frank Koskey, III, Matthew George White, Javier Becerra
August 05, 2025

As the race for real-time data access intensifies, organizations are confronting a growing legal and operational challenge: web scraping. What began as a fringe tactic by hobbyists has evolved into a sophisticated, multibillion-dollar ecosystem driven by commercial data aggregators. Think of it as digital fly fishing – automated bots cast wide nets across public-facing websites, harvesting pricing data, product listings, reviews, and more, often faster than a human could click "refresh." These entities now routinely circumvent traditional barriers to access – not by breaching platforms directly, but by piggybacking on legitimate users' access to bypass technical and contractual restrictions.

Understanding the mechanics of web scraping as well as how aggregators exploit contractual workarounds is important for businesses and organizations so that they can regain control of their data through thoughtfully structured agreements and properly implemented and configured technology, particularly with respect to APIs ("Application Programming Interface") and direct data licensing.

How Are Data Aggregators Obtaining Your Data?

Web Scraping

Web scraping refers to the use of automated tools or programs to collect large amounts of data from websites or internet-based applications. Rather than browsing a website like a human would, these tools pull information directly from the website's code or data feeds and harvest large volumes of structured data. The data sought by web scrapers includes things like prices, product listings, user reviews, public records, and transactional histories.

While scraping isn't inherently malicious, in fact, it can even serve legitimate purposes, such as powering academic research projects, digital archiving, or competitive benchmarking. However, when used for commercial purposes, it raises both legal and ethical concerns. Unauthorized scraping may violate terms of service, exceed authorized access under laws like the Computer Fraud and Abuse Act (CFAA), or infringe on intellectual property rights. Beyond the legal risk, scraping can strain servers, distort website analytics, and erode a business's ability to control or commercialize its own information. What starts as a technical workaround can quickly become a business and legal flashpoint.

Consider a recent example where web scraping is taking center stage in a legal battle:

***Reddit, Inc. v. Anthropic, PBC*, 2025 WL 1617126 (Cal. Super. Ct. 2025).**

Reddit recently filed a complaint against Anthropic, the developer of the Claude AI model, alleging that Anthropic scraped Reddit's platform without permission and used that data to train and commercialize Claude. Reddit claims the scraping violated its User Agreement, which prohibits commercial use and scraping without a license. Reddit brings claims for breach of contract, trespass to chattels, tortious interference, and unfair competition.

End-User Consent

In response to lawsuits and public backlash, many large-scale data aggregators now avoid direct scraping. Instead, they exploit through a subtler approach: contracting directly with a platform's end users, often individual account holders, and asking them to provide access to their accounts.

For example, a financial aggregator might ask bank customers to "link their account" by logging into their online banking interface. Once linked, the aggregator collects transaction history, balances, or other account data, either by scraping the site using the customer's credentials or through an authorized API connection. Even though the platform itself (the bank, in this case) never gave permission, the aggregator's access is arguably lawful because the customer agreed.

This workaround allows aggregators to sidestep many direct enforcement tools. Because aggregators don't hack into a platform's systems, traditional cybersecurity laws such as the CFAA may offer little recourse. Instead, they rely on the guise of user consent, leveraging the customer's access to do what the aggregator itself couldn't do directly.

Why This Matters: Risks to Platforms and Data Hosts

Unauthorized (and often unchecked) web scraping and end-user access workarounds can do serious damage to the platforms hosting the data, such as:

- **Loss of Control:** Aggregators dictate how the data is stored, used, and monetized. Platforms lose control over how their proprietary or sensitive data is distributed, reformatted, or resold. If an organization relies on data as its revenue source, the copying and reuse of data by aggregators undermines the organization's business model and devalues the content.

- **Security Risks and Operational Costs:** Credential sharing (especially when aggregators use scraping rather than API access) creates cybersecurity vulnerabilities and increases the risk of breaches or unauthorized transactions. It can also lead to increased operational costs, with excessive loads on servers and potentially slowing down performance for legitimate users.

- **Brand and Trust Erosion:** If an aggregator misuses data or suffers a breach, customers often blame the originating platform, even if it is not involved.

- **Regulatory Exposure:** In industries like finance, health care, or insurance, platforms may face compliance risks if customer data is accessed or transmitted in ways that violate privacy laws, even indirectly.

The Solution: Take Control Through API Agreements and Direct Licensing

Rather than playing whack-a-mole with each aggregator individually or trying (and often failing) to block scraping entirely, many platforms are taking a more strategic and proactive approach, channeling access through an API agreement. These agreements offer a secure, structured gateway that allows third parties to access the specific data fields *under defined conditions* with guardrails for security, usage, and compliance baked in. Platforms can pair API access with data use agreements that:

- Specify permitted uses and storage limitations
- Require regular security audits and data retention practices
- Prohibit sublicensing or resale of data
- Include indemnity and enforcement provisions
- Allow for termination if terms are violated

By allowing aggregators to contract directly with the platforms rather than through end-user consent, platforms can impose restrictions, track data usage, and avoid downstream risks of scraping or shadow access.

Conclusion and Best Practices

If your business or organization hosts valuable or sensitive user data, you're likely already in the crosshairs of commercial data aggregators – whether you know it or not. Even with anti-scraping measures in place, aggregators are brazenly exploiting indirect access channels, like end-user credentials, to harvest your data at scale.

This isn't just a technical issue, it's a business risk, a legal liability, and a threat to data governance. Therefore, proactive action is needed. Consider implementing best practices to help mitigate the risk of commercial web scraping:

1. **Strengthen Terms of Use**: Review your terms of service and data-sharing policies to ensure they clearly prohibit unauthorized scraping and downstream use. Additionally, businesses and organizations should ensure that users affirmatively accept such terms.

2. **Assess Access Controls and Use Technical Barriers**: Evaluate how users can share or delegate access, and whether that access effectively circumvents your platform's controls. Furthermore, consider technical measures to make it more difficult for web scrapers to access your data at scale, including rate limiting to prevent high-volume requests, bot detection tools to analyze traffic patterns, and CAPTCHAs to distinguish human users from bots.

3. **Control Potential Data Exposure**: Consider API licensing models that provide structured access while preserving your platform's security, business model, and legal rights. This includes limiting the availability of high-value data, avoiding exposure through unauthenticated APIs, and delaying the loading of critical content when appropriate.

4. **Proactive Enforcement**: Consult legal counsel about your potential options for recourse when scraping is detected. This may include cease-and-desist letters, DMCA takedown notices if your content is reposted, breach of contract claims, CFAA claims, and unfair competition or misappropriation doctrines.

At Baker Donelson, we are actively helping clients stay ahead of the curve where data, platform security, and legal strategy intersect. Our multi-disciplinary teams have deep experience advising clients on platform security, API agreements, data protection frameworks, and enforcement strategies across sectors and industries where the stakes are high and the data is valuable. Whether you need to tighten your terms, build a secure API framework, or take legal action, we can guide you every step of the way. If you have questions about your data strategy or how to protect against unauthorized data access, reach out to any of our authors, Alexander F. Koskey, CIPP/US, CIPP/E, PCIP, Matthew G. White, AIGP, CIPP/US, CIPP/E, CIPT, CIPM, PCIP, and Javier Becerra, CIPP/US, or any member of our Data Protection, Privacy and Cybersecurity Team.

Kaytlyn Mullins, a summer associate at Baker Donelson, contributed to this article.