NewtonX

WHITEPAPER

Data Extraction Prevention Best Practices Study

The why: data is key to the digital economy

Data drives the modern digital economy, but with massive amounts of data available on the internet comes massive potential to misuse this data. One form of data misuse is data scraping: an automated way of accessing and extracting data. While there are legitimate use cases, data scraping can be harmful in the wrong hands for users and organizations alike. Data scraping can harm user privacy and increase the liability and reputational risk for the companies holding the data.

Prevention of malicious data scraping is possible with the proper procedures, stakeholder involvement, and collaboration. NewtonX tapped the expertise of more than 1,300 professionals in the field across industries to understand the current state of data scraping, data scraping prevention, and best practices. These best practices along with increased awareness and ongoing cross-organizational conversations can help us all be better prepared against the dark side of data scraping.

The research methodology: tapping 1,300+ experts with experience in scraping detection & prevention

Custom recruiting. In order to better understand best practices in data access and anti-scraping, we custom recruited more than 1,300 professionals with experience in scraping detection and prevention. We found experts from US- and Europe-based companies with more than 50 employees and across industries: social media, commerce, gig economy, online dating, gaming, financial services, and IT services. We recruited experts who were not only involved in scraping detection and prevention, but also accountable if their organization's user data were scraped.

Identity verification. Market research conducted in the wrong way can be riddled with fraud, so we passed all participants through our standard 2-point identity verification process. This way, we knew that respondents were who they said they were and could trust that we were getting high quality insights.

Methodology. We used our Q3 approach to capture both breadth and depth for this study. We started with twenty-four 60-minute qualitative interviews, moved into web surveys with 1,300 experts, then closed out with fourteen individual 60-minute qualitative follow-up interviews. This allowed us to thoughtfully craft the survey design, scale up the research, and do a deep-dive on findings to further refine insights and best practices.

The findings: anti-data scraping awareness and efforts are varied

Through our research we found a range of perspectives and preparedness around data extraction and scraping prevention. Some organizations have dedicated strategies and resources around data scraping prevention, while others perceive data scraping as less of a priority or perhaps not even negatively impactful. This variation could stem from different definitions of data, awareness of scraping impact, stakeholder alignment, or ability to exchange information, but one thing is clear: there is room for improvement. Our key insights are detailed below.

User data is important to protect—but how and to what extent? We all understand the significance of user data in the digital era. We use user data to build businesses, to improve user experiences, and to create virtual spaces where people can interact and exchange ideas, goods, and services. 52% of our surveyed experts stated user data is the most important type of data for their organization.

87% **of experts stated that user data scraping prevention is important or very important relative to other security issues.**

Even though the importance of user data and its protection is common knowledge, the definition of user data varies across organizations and can encompass a lot. Thus, the amount of user data that is protected similarly varies. Some organizations may focus on guarding only the most sensitive personally identifiable information (think social security numbers, driver's license numbers, and biometrics), while others adopt a broader view of personal data protection.

GDPR maintains a broad definition: personal data is "any information relating to an identified or identifiable natural person." As such, it's important for us all to look beyond the most sensitive types of personal data and expand our idea of user data and what needs to be protected. When we widen that lens, we can see all the opportunities for data misuse.

The impact and extent of data scraping may be hidden. In our research, we found that many experts don't perceive that users are negatively impacted by data scraping. The potential for data scraping is a necessary consequence of an open internet; as such, it can be tricky to draw a line between what kind of data scraping is "good" versus "bad." Even if an organization acknowledges the downsides of unbridled data scraping, it may be willing to accept a certain level of risk in order to minimize prevention techniques, which some see as causing "friction" for users.

The challenge here is that the impact of data scraping can be subtle, but pervasive. Data scraping happens without many users knowing; it has minimal direct impact to users in a way that's easily visible to them. Similarly, data scraping misuse may go unnoticed. This is in part because of data scraping's more nuanced nature: it isn't definitively negative the way that data breaches and hacks are. Perhaps it's also because data scraping is generally a persistent activity behind the scenes that doesn't call a lot of attention unless something goes majorly wrong. Some experts admit their organizations may not always be aware of the extent of data scraping or what the scraped data is being used for. At the same time, 89% of respondents have had user data scraped. Clearly, user data is commonly scraped, but it happens quietly and continuously without drawing much attention or action. And if we're not paying attention, we may also be missing the instances of data scraping misuse. As one head of data in media shared, **"from my view as a data person, we are not rating [data scraping] as important enough."**

89% **of respondents have had user data scraped.**

The good news is there are data scraping prevention measures available. Additionally, for those concerned with adding friction for users, data scraping prevention and user experiences don't have to be at odds with one another. In fact, many experts don't perceive scraping prevention measures as negatively impacting users. Users have grown accustomed to scraping prevention methods like CAPTCHAs, and in some cases, such methods actually help create peace of mind about site security. Just as data scraping goes mostly unnoticed by users, data scraping prevention measures can go mostly unnoticed too.

Data scraping prevention requires data classification. It's hard to plan against risks without first understanding what exactly those risks are. When it comes to crafting data scraping prevention measures, organizations need to understand what they're working with and what they're up against. This means understanding what data is being collected, where it is stored, how it is exposed, and who has access. By classifying data in this way, organizations can then prioritize and rank the sensitivity of the data, which then helps quantify the impact of scraping. This quantified impact can then spur awareness around data scraping prevention needs within an organization—because what gets measured gets managed.

Anti-data scraping efforts require processes, policies, and procedures across departments. Once an organization understands the current impact of data scraping, the next step is to work cross-departmentally to develop strategies and standards. Our research indicates that implementing processes, policies, and procedures related to data scraping management and incidents is important—70% of respondents stated they have these in place. Surprisingly, only 42% of respondents have a dedicated strategy to deal with data scraping. This mismatch suggests that while most organizations have processes, policies, and procedures to handle incidents, more could be done to develop longer term plans and goals around data scraping.

42% **of respondents have a dedicated strategy to deal with data scraping.**

When it comes to designating who is responsible for data scraping management and incidents, many of our respondents have anti-data scraping teams that report into two or more departments—most commonly some combination of IT or information security, data management, or legal/compliance. This appears to be a best practice, and as one data engineering manager at a large retailer shared, **"Stakeholders like pricing, legal team, compliance, IT and whoever else; it is important everyone is involved and aware."**

After teams and processes have been put in place, it is essential to document them, reflect, and iterate. Reviewing incidents and procedures is one important way to improve data scraping prevention. As an engineering manager in the travel industry shares, **"scraping is constantly evolving and so we have to evolve our prevention as well. Retrospective reviews help us to learn and evolve."** Additionally, blending human judgment and automation can help organizations continuously improve techniques for detection and prevention at scale—86% of respondents shared they use some combination of automation and human involvement. A security engineer at a technology company pointed out, **"setting and forgetting doesn't work. It requires human judgment."**

Anti-data scraping efforts would benefit from knowledge sharing. For those who don't have the time or resources to tackle data scraping detection and prevention on your own, you're not alone. Relying on outside help is underlined as a best practice and 64% of respondents rely on external service providers.

Another method to improve anti-data scraping techniques is to have active exchanges within and between organizations. This intentional sharing of experiences can improve internal alignment around data scraping prevention and foster more innovative and effective ideas across organizations. Respondents indicated exchanges with external peers are currently limited and that a formalized forum would offer participants the opportunity to learn from each other and benefit. As a compliance manager at a utilities organization expressed, **"there is a lack of community around this topic. We are stronger together and [community] would probably help me avoid reinventing the wheel."**

The best practices: guidance for data extraction prevention

Our research highlighted a need to increase awareness around the importance of data extraction prevention, to have more conversations on this topic—both within and across organizations, and to establish standards around anti-data scraping efforts. With that in mind, we've synthesized the research into 11 best practices that management and security stakeholders can use to shape data extraction detection and prevention measures for their organizations. To dive deeper into our findings, view our full study here.

1. **MAP YOUR ORGANIZATION'S DATA**
   Map the types of data your organization processes and where it resides. Grade the data's sensitivity, the impact if it were to be scraped, and the priority it has for the business.

2. **DETERMINE DATA ACCESS REQUIREMENTS & EXPOSURE**
   Document who internally (business users) and externally (users on site, partners, aggregators, etc.) needs access to what data and how it is being exposed.

3. **DISCOVER WHAT IMPACT DATA SCRAPING HAS**
   Review traffic data, look externally at how other organizations are impacted by scraping, and quantify the magnitude of data extraction.

4. **GAIN INTERNAL EXECUTIVE SPONSORSHIP & AWARENESS**
   Use the quantified impact to raise awareness with senior leadership and gain sponsorship to build an anti-data scraping strategy.

5. **ESTABLISH MAIN STAKEHOLDERS AND RESPONSIBILITIES**
   Build a team of stakeholders from different departments and determine accountability and responsibilities.

6 **DETERMINE WHAT EXTERNAL EXPERTISE IS NEEDED**
With the committee of internal stakeholders, determine gaps in expertise and knowledge and fill them with external services.

7 **ESTABLISH DETECTION METHODS ACROSS PRODUCTS**
Build a robust monitoring system that collects data on traffic at every point in the product architecture.

8 **DESIGN YOUR PREVENTION TECHNIQUE MIX**
Consider the level of acceptable risk along with the balance between prevention and user experience, and build a technique mix.

9 **ESTABLISH PROCEDURES FOR REVIEW AND IMPROVEMENT**
Establish procedures for reviewing incidents and data. Create a culture of continuous improvement in the security posture.

10 **CREATE AND SHARE DOCUMENTATION FROM THE OUTSET**
Ensure all procedures, policies and choices are documented and shared internally with the right stakeholders.

11 **SHARE BEST PRACTICES WITH EXTERNAL ORGANIZATIONS/PEERS**
Learn from other organizations on how to improve techniques and avoid reinventing the wheel.

About NewtonX Current:

NewtonX Current is the proprietary research arm of NewtonX. Current research is designed, coded and fielded by our team of senior researchers to answer today's most pressing questions. Each publication will be informed by topics of interest from the B2B research community and will range from trends in Cybersecurity to Advertiser spending. We're excited to share B2B insights on the topics that matter most.

This study was sponsored by Meta and done in collaboration with 1,300 anti-data scraping experts from varied industries. Meta had no influence on the outcomes of the study.

NewtonX

As the world's leading B2B research company, NewtonX fields large-scale quantitative surveys, facilitates qualitative interviews, engages in long-term consultations, and creates customized research plans.

We partner with the world's top consultancies, marketers, and technology companies. Together with our clients, we're ushering in a new standard of truth in B2B research.

To learn more, visit newtonx.com/get-in-touch