Joint statement on data scraping and the protection of privacy

August 24, 2023

Key takeaways

- **Personal information that is publicly accessible is still subject to data protection and privacy laws in most jurisdictions.**
- **Social media companies and the operators of websites that host publicly accessible personal data have obligations under data protection and privacy laws to protect personal information on their platforms from unlawful data scraping.**
- **Mass data scraping incidents that harvest personal information can constitute reportable data breaches in many jurisdictions.**
- **Individuals can also take steps to protect their personal information from data scraping, and social media companies have a role to play in enabling users to engage with their services in a privacy protective manner**

Introduction

1. Data scraping generally involves the automated extraction of data from the web. Data protection authorities are seeing increasing incidents involving data scraping, particularly from social media and other websites that host publicly accessible data.
2. The capacity of data scraping technologies to collect and process vast amounts of individuals' personal information from the internet raises significant privacy concerns, even when the information being scraped is publicly accessible.
3. In most jurisdictions, personal information that is "publicly available", "publicly accessible" or "of a public nature" on the internet, is subject to data protection and privacy laws. Individuals and companies that scrape such personal information are therefore responsible for ensuring that they comply with these and other applicable laws. However, social media companies and the operators of other websites that host publicly accessible personal information (SMCs and other websites) also have data protection obligations with respect to third-party scraping from their sites. These obligations will generally apply to personal information whether that information is publicly accessible or not. Mass data scraping of personal information can constitute a reportable data breach in many jurisdictions.
4. Scraped personal information can be exploited for various purposes, such as monetization through re-use on third-party websites, sale to malicious actors, or private analysis or intelligence gathering, resulting in serious risks to individuals as explained further below.
5. SMCs and other websites should carefully consider the legality of different types of data scraping in the jurisdictions applicable to them and implement measures to protect against unlawful data scraping.

6. **The aim of this joint statement is to**:
    - Outline the key **privacy risks** associated with data scraping;
    - Set out **how SMCs and other websites should protect individuals' personal information** from unlawful data scraping to meet regulatory expectations; and
    - Set out **steps that individuals can take** to minimise the privacy risks from scraping.
7. We have published this joint statement for the benefit of SMCs and other websites, as well as for individuals who use and post personal information on these websites. It has also been sent directly to Alphabet Inc. (YouTube), ByteDance Ltd (TikTok), Meta Platforms, Inc. (Instagram, Facebook and Threads), Microsoft Corporation (LinkedIn), Sina Corp (Weibo), and X Corp. (X, previously Twitter).
8. The practices outlined in this joint statement reflect common global data protection principles and practices, and are designed to help protect against data scraping of personal information and mitigate against its privacy impacts. While the expectations are phrased as recommendations (using the term "should"), many of them are explicit statutory requirements in particular jurisdictions or may be interpreted as such by courts and data protection authorities.
9. We recognise that some SMCs have implemented controls to address data scraping of publicly accessible personal information, including for example, through court action or governance initiatives. The principles and expectations included in this open letter are informed by, and build on, that activity.

Privacy risks

10. In recent years, many data protection authorities have seen increased reports of mass data scraping from SMCs and other websites. The reports raise a number of privacy concerns, including the use of scraped data for:
    - **Targeted cyberattacks** – for example, scraped identity and contact information posted on 'hacking forums' may be used by malicious actors in targeted social engineering or phishing attacks.
    - **Identity fraud** – scraped data may be used to submit fraudulent loan or credit card applications, or to impersonate the individual by creating fake social media accounts.
    - **Monitoring, profiling and surveilling individuals** – scraped data may be used to populate facial recognition databases and provide unauthorised access to authorities.
    - **Unauthorised political or intelligence gathering purposes** – scraped data may be used by foreign governments or intelligence agencies for unauthorised purposes.
    - **Unwanted direct marketing or spam** – scraped data may include contact information that can be used to send bulk unsolicited marketing messages.
11. More broadly, individuals lose control of their personal information when it is scraped without their knowledge and against their expectations. For example, data scrapers may aggregate and combine scraped data from one site with other personal information, and use it for unexpected purposes. This can undermine individuals' trust in the SMC or other websites, with potentially detrimental impacts on the digital economy. Moreover, even if individuals decide to delete their information from a social media account, data scrapers will likely continue using and sharing

information they have already scraped, limiting individuals' control over their online presence and reputation.

SMCs and other websites should protect personal information from unlawful data scraping

12. **SMCs and other websites are responsible for protecting individuals' personal information from unlawful data scraping**.

13. Techniques for scraping and extracting value from publicly accessible data are constantly emerging and evolving. **Data security is a dynamic responsibility** and vigilance is paramount.

14. As no one safeguard will adequately protect against all potential privacy harms associated with data scraping, SMCs and other websites should implement **multi-layered technical and procedural controls to mitigate the risks**. A combination of these controls should be used that is proportionate to the sensitivity of the information, and may include:

    - Designating a team and/or specific roles within the organisation to identify and implement controls to protect against, monitor for, and respond to scraping activities.
    - 'Rate limiting' the number of visits per hour or day by one account to other account profiles, and limiting access if unusual activity is detected.
    - Monitoring how quickly and aggressively a new account starts looking for other users. If abnormally high activity is detected, this could be indicative of unacceptable usage.
    - Taking steps to detect scrapers by identifying patterns in 'bot'[1] activity. For example, a group of suspicious IP addresses can be detected by monitoring from where a platform is being accessed by using the same credentials from multiple locations. This would be suspicious where these accesses are occurring within a short period of time.
    - Taking steps to detect bots, such as by using CAPTCHAs[2], and blocking the IP address where data scraping activity is identified.
    - Where data scraping is suspected and/or confirmed, taking appropriate legal action such as the sending of 'cease and desist' letters, requiring the deletion of scraped information, obtaining confirmation of the deletion, and other legal action to enforce terms and conditions prohibiting data scraping.
    - In jurisdictions where the data scraping may constitute a data breach, notifying affected individuals and privacy regulators as required.

[1] A 'bot' – a computer program that performs automatic repetitive tasks, or a computer application designed to automate certain tasks (such as gathering information online), especially one designed to perform a malicious action. Merriam-Webster Dictionary

[2] A CAPTCHA is a **C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part. This is a program that tests whether a user is a human or an automated program (e.g. a bot) (PC Mag, *Definition of CAPTCHA*). Some examples of CAPTCHAs are programs that require a user to: interpret text that is distorted, or look at a set of similar pictures and identify which of these contain a specific object.

15. In addition to security controls like those mentioned above, SMCs and other websites also have a role to play in enabling users to engage with their services in a privacy protective manner. To this end, SMCs and other websites should proactively support their users so that they can make informed decisions about how they use the platform and what personal information they share. This should also involve increasing user awareness and understanding of the privacy settings they can utilize, as discussed further below.

16. If any safeguards implemented to protect against data scraping involve processing of personal information, SMCs and other websites should ensure that this processing complies with any applicable data protection or privacy law requirements. As a matter of good practice and to ensure transparency, these entities should also inform their users of the steps they have taken to protect against data scraping.

17. Given the dynamic nature of data scraping threats, SMCs and other websites should continuously monitor for, and respond with agility to, new security risks and threats from malicious or other unauthorised actors to their platform. Controls should be routinely stress-tested and updated to ensure that they remain effective and keep pace with changing technologies. SMCs and other websites should also collect and analyse metrics on scraping incidents, to inform and identify areas of improvement in their security control framework.

Steps that individuals can take to minimise the privacy risks from data scraping

18. Although the security controls outlined above may mitigate the risks associated with data scraping, no safeguards are 100% effective and individuals should therefore be mindful that the personal information they share online may be at risk.

19. While this joint statement focuses on the measures that SMCs and other websites can implement to mitigate against the risk of data scraping, individuals can also take steps to empower themselves and better protect their personal information, including:

    - **Read the information provided by the SMC or other website about how they share personal information, including the privacy policy** – Specifically focussing on the website's policies on sharing and disclosure will assist individuals in making an informed decision on what information they choose to share, and in understanding the resulting privacy risks.
    - **Think about the amount and kinds of information shared** – Individuals should consider limiting the information that they post online. In particular, individuals should be cautious to limit the sharing of sensitive information and consider if sharing certain information (such as personal details, account numbers or identification numbers) may put them at risk of reputational damage, discrimination, harassment, identity fraud or theft.
    - **Understand and manage privacy settings** – While individual-user privacy settings can only go so far in providing privacy protection, they can and should help individuals increase the control they have over how their personal information is shared online. Accordingly, website users should consider using these settings to limit the information that they make publicly accessible.

20. Ultimately, we encourage individuals to **think long term**. How would a person feel years later, about the information that they share today? While SMCs and other websites may offer tools to

delete or hide information, that same information can live forever on the Web if it has been indexed or scraped, and onward shared.

21. If individuals are concerned that their data may have been scraped unlawfully, or improperly, then they can contact the SMC or website, and if dissatisfied with the response, they can file a complaint with their relevant data protection authority. They may also wish to review their privacy settings and the information that they are sharing online, to make changes and remove personal information as needed.

Conclusion

22. The expectations in this joint statement set out key areas for SMCs and other websites to focus on with a view to ensuring that they protect personal information accessible on their websites from data scraping, particularly so that they are compliant with data protection and privacy laws around the world. Protecting against data scraping will also support SMCs and other websites in building the trust and confidence of their userbase.

23. SMCs and other websites can further protect their users' information and reinforce user trust by actively informing their users of the steps they can take to protect their personal information, like those outlined above.

24. We welcome any feedback from SMCs by 1 month from the issuance of this statement demonstrating how they comply with the expectations outlined in this joint statement. Any responses will be shared amongst signatories and may be published.

This statement is endorsed by the following members of the GPA's International Enforcement Cooperation Working Group ("IEWG").

Elizabeth Hampton
Deputy Commissioner
Office of the Australian Information Commissioner
Australia

Philippe Dufresne
Commissioner
Office of the Privacy Commissioner of Canada
Canada

Stephen Bonner
Deputy Commissioner – Regulatory Supervision
Information Commissioner's Office
United Kingdom

Ada CHUNG Lai-ling
Privacy Commissioner
Office of the Privacy Commissioner for Personal Data
Hong Kong
China

Adrian Lobsiger
Commissioner
Federal Data Protection and Information Commissioner
Switzerland

Tobias Judin
Head of International Section
Datatilsynet
Norway

Michael Webster
Privacy Commissioner
Office of the Privacy Commissioner
New Zealand

Cielo Angela Peña Rodriguez
Deputy Superintendent for the Protection of Personal Data
Superintendencia de Industria y Comercio
Colombia

Paul Vane
Information Commissioner
Jersey Office of the Information Commissioner
Jersey

Omar Seghrouchni
President
CNDP (Commission Nationale de contrôle de la protection des Données à caractère Personnel)
Morocco

Beatriz de Anchorena
Director
AAIP (Agency for Access to Public Information)
Argentina

Josefina Román Vergara
Commissioner
National Institute for Transparency, Access to Information and Personal Data Protection (INAI)
Mexico