Concluding joint statement on data scraping and the protection of privacy

Informed by engagement with industry on the initial Joint Statement on Data Scraping and the Protection of Privacy (August 2023)

October 2024

Key takeaways

Initial Statement

This Concluding Statement builds on the Joint statement on data scraping and the protection of privacy (the Initial Statement), published August 24, 2023, which highlighted the following key messages:

- Personal information that is publicly accessible is subject to data protection and privacy laws in most jurisdictions.
- Social media companies (SMCs) and the operators of websites that host publicly accessible personal data have an obligation to protect publicly accessible personal data on their platforms from data scraping that violates data protection and privacy laws ("unlawful 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.

Concluding Statement

Based on engagement with SMCs and other industry stakeholders that followed the issuance of the Initial Statement, the co-signatories wish to highlight the following additional key takeaways:

- To effectively protect against unlawful scraping, organizations should deploy a combination of safeguarding measures, and those measures should be regularly reviewed and updated to keep pace with advances in scraping techniques and technologies.
- While artificial intelligence (AI) is used by some sophisticated data scrapers to evade detection, it can also represent part of the solution, serving to enhance protections against unlawful scraping.
- The obligation to protect against unlawful scraping applies to both large corporations and Small and Medium Enterprises (SMEs). There are lower-cost measures that SMEs can implement, with assistance from service providers, to meet this obligation.

- Where SMCs and other organizations contractually-authorize scraping of personal data from their platforms, those contractual terms cannot, in and of themselves, render such scraping lawful; however, they can be an important safeguard.
    - Organizations who permit scraping of personal data for any purpose, including commercial and socially beneficial purposes, must ensure without limitation, that they have a lawful basis for doing so, are transparent about the scraping they allow, and obtain consent where required by law.
    - Organizations should also implement adequate measures, including contractual terms and associated monitoring and enforcement, to ensure that the contractually authorized use of scraped personal data is compliant with applicable data protection and privacy laws.
- When an organization grants lawful permission for third parties to collect publicly accessible personal data from its platform, providing such access via an Application Programming Interface (API)[1] can allow the organization greater control over the data, and facilitate the detection and mitigation of unauthorized scraping.
- SMCs and other organizations that use scraped data sets and/or use data from their own platforms to train AI, such as Large Language Models, must comply with data protection and privacy laws as well as any AI-specific laws where those exist. Where regulators have made available guidelines and principles on the development and implementation of AI models, we expect organizations to follow that guidance.

Introduction

1. The initial Joint Statement on data-scraping and the protection of privacy (the Initial Statement), published in August 2023, set out expectations regarding what organizations should do to ensure that individuals are protected from the risks resulting from unlawful scraping. The present Concluding Statement was developed to reinforce the requirements set out in the Initial Statement, share best practices and lessons learned through engagements with SMCs and industry stakeholders following the publication of that statement, and set out further expectations for SMCs and other organizations that host publicly accessible personal information.

[1] Application Programming Interface (API) - a way of communicating with a particular computer program or internet service.

2. Both statements address data scraping in the form of automated extraction of personal data from the web. These statements do not address indexing by search engines, nor do they address the scraping of non-personal information.
3. While the Initial Statement was published by 12 members of the International Enforcement Working Group (IEWG) and endorsed by two additional members following its publication, the Initial Statement and this Concluding Statement are now endorsed by a total of 16 co-signatories[2].

Engagement with industry

4. After issuing the Initial Statement, the co-signatories shared a copy with 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) inviting them to comment on how they comply with the expectations outlined in the document.
5. Over the course of the following months, the co-signatories engaged with several of these organizations, in writing and through virtual engagements. The co-signatories also engaged with the Mitigating Unauthorized Scraping Alliance (MUSA), which approached the co-signatories to share its perspectives on mitigation against unauthorized scraping.[3]
6. The co-signatories were also approached by a commercial data scraping company that shared details regarding its efforts towards lawful collection of publicly accessible data (which can include personal data). While this Concluding Statement, and the Initial Statement, are not primarily directed at data scrapers, commercial data scrapers should take note that publicly accessible personal data will generally be subject to data protection and privacy laws, and as such, they should implement measures to comply with those laws.
7. Through these exchanges, the co-signatories were able to engage with industry meaningfully, in a coordinated manner and with a unified voice. In turn, this provided relevant stakeholders with the opportunity to explain their respective

---

[2] Office of the Australian Information Commissioner (OAIC); Office of the Privacy Commissioner of Canada (OPC-Canada); United Kingdom Information Commissioner's Office, (ICO); Hong Kong Office of the Privacy Commissioner for Personal Data (PCPD); Norway Data Protection Authority (Datatilsynet); Swiss Federal Data Protection and Information Commissioner (FDPIC); Colombian Superintendencia Industria y Comercio (SIC); Office of the Privacy Commissioner of New Zealand (OPC-New Zealand); Jersey Office of the Information Commissioner (JOIC); Moroccan Commission Nationale de Contrôle de la Protection des Données à Caractère Personnel (CNDP); Argentine Agencia de Acceso a la Información Pública (AAIP); Mexican Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales, (INAI); Guernsey Office of the Data Protection Authority (ODPA); Spain Agencia Española de Protección de Datos (AEPD); Monaco Commission de Contrôle des Informations Nominatives (CCIN); Israel Privacy Protection Authority (PPA).

[3] The Mitigating Unauthorized Scraping Alliance describes itself as an organization that unites industry and regulators to combat unauthorized data scraping, aiming to promote best practices, raise public awareness, and provide valuable insights to policymakers.

approaches to data and privacy protection, through direct and practical interactions with a diverse subset of the global privacy regulatory community.

8. Below, the co-signatories share lessons learned from their discussions with industry representatives, as well as additional expectations for organizations that host publicly accessible personal data.

Lessons learned and co-signatories' expectations

9. As with the Initial Statement, many of the recommendations below represent statutory requirements in some or all jurisdictions.
10. A fundamental takeaway from the Initial Statement is that publicly accessible personal data is still subject to data protection and privacy laws in most jurisdictions. SMCs and 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 scraping.

Challenges and solutions in keeping up with advances in data scraping practices

11. In the Initial Statement, the co-signatories highlighted the need for SMCs and other organizations to implement a multi-layered approach to protecting publicly accessible data on their platforms from unlawful scraping.
12. Through our engagements that followed the issuance of that statement, we established that, while SMCs face challenges in protecting against unlawful scraping (such as increasingly sophisticated scrapers, ever-evolving advances in scraping technology, difficulty in differentiating scrapers from authorized/lawful users, and the need to maintain a user-friendly interface), they are motivated to protect against unauthorized scraping.
13. SMCs generally confirmed that they have implemented many of the measures identified in the Initial Statement, such as, and without limitation:
    - Designating a team and/or specific roles within the organization to develop 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.
    - Taking steps to detect scrapers and "bot"[4] activity, such as using CAPTCHAs[5] and blocking IP addresses where such activity is identified.

[4] A bot is an automated software application that performs repetitive tasks over a network. It can follows specific instructions to imitate human behavior.

[5] A CAPTCHA is a program or system intended to distinguish human from machine input.

- Where data scraping is suspected and/or confirmed, taking appropriate legal action, such as sending "cease and desist" letters, requiring the deletion of scraped information, and obtaining confirmation of the deletion.
- Closely monitoring the threat landscape and new technologies to develop and adjust safeguards accordingly.

14. Through our engagements, we also learned of further measures, beyond those detailed in the Initial Statement, that organizations employ to protect against data scraping, such as the implementation of platform design elements that make it harder to scrape data using automation (e.g., random account URLs, random interface design elements, and tools to detect and block malicious internet traffic).

15. We learned that the rapid emergence of AI can represent a threat to privacy. SMCs told us that scrapers are now using AI to scrape data more effectively (e.g., via "intelligent" bots that can simulate real user activity). At the same time, SMCs explained that they too are employing AI to better detect and protect against unauthorized scraping, highlighting that innovative AI tools can also be part of the solution.

16. Ultimately, the co-signatories learned that while no measure is guaranteed to protect against all unlawful scraping - since sophisticated low-volume scraping can often resemble user activity - a multi-layered and dynamic combination of safeguards can be particularly effective in protecting against mass scraping and the amplified harms that can result when a large volume of data subjects are affected.

Small and medium enterprises (SMEs)

17. SMEs rarely have the same financial resources or technical capabilities as global SMCs. This does not, however, absolve SMEs of their responsibility to protect against unlawful scraping. Indeed, many SMEs host large amounts of publicly accessible personal data, which should be protected by a multi-layered combination of technical and procedural controls against data scraping.
18. The co-signatories learned from their engagement with industry that there is a variety of tools available to protect against unlawful scraping. Some of those tools, such as bot detection, rate limiting and CAPTCHAs, can be accessible to SMEs on a more modest budget. There are also third-party service providers who can assist SMEs in protecting against unlawful scraping. However, the co-signatories wish to emphasise that engaging a third-party service provider does not absolve the organization of its own responsibility to protect personal data.
19. Ultimately, under data protection and privacy laws, safeguards should be appropriate and commensurate to the sensitivity of the information in question. Organizations should therefore limit the amount and sensitivity of information they

make publicly accessible to that which they can adequately protect from unlawful scraping.

SMC-allowed scraping and lawful scraping

20. Several SMCs indicated that in certain circumstances, they allow scraping or other forms of mass collection of data from their platforms (e.g., through API access, discussed further below), in furtherance of their own or third parties' commercial interests, such as those associated with platform management.
21. The companies explained that they generally "authorize" such collection via contractual terms, such as those in their Terms and Conditions. SMCs further explained that to ensure that the scraping that they permit is lawful, their contractual terms generally require third parties on their platform to comply with applicable laws. They also explained that it can be difficult for them to determine whether scraped data is used by those parties solely for purposes allowed by their contract.
22. The co-signatories note that contractual terms cannot in and of themselves render data scraping lawful. For example, organizations must also ensure that they have a lawful basis for granting access or permitting collection of personal data, that they are transparent about the scraping they allow, and that they obtain consent where required by law.
23. Furthermore, while contractual terms are an important safeguard against unlawful scraping, a contractual term indicating that third parties must comply with applicable laws is not sufficient. Organizations should implement adequate measures to ensure that contractually-allowed use of scraped personal data is compliant with applicable data protection and privacy laws. The contract could, for example, specify limitations on the information that may be scraped and the purposes for which it may be used, as well as the consequences for non-compliance with those terms. However, organizations cannot simply rely on contractual measures. They should also implement measures to monitor third parties' compliance with contractual limitations, and to enforce compliance when those terms are not respected.

Access to data for research and other potentially socially beneficial purposes

24. In certain circumstances, SMCs may be required by law to provide third parties, such as researchers, with large-scale access to publicly accessible data on their platforms (e.g., pursuant to Article 40 of the EU Digital Services Act[6]). In other

[6] Article 40, Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act):

circumstances, we learned that SMCs may choose to provide data access to third parties, even where there is no legal requirement to do so (e.g., in support of socially beneficial research). Several of the companies indicated that they often provide such access via an API, in particular where they are required or permitted by law to grant large-scale access.

25. While the co-signatories acknowledge the importance of socially beneficial research, they wish to remind SMCs and other organizations that host publicly accessible personal data that, when allowing large-scale access or collection, organizations must ensure that they are complying with applicable data protection and privacy laws, including by ensuring that there is a lawful basis for granting access or permitting collection. Specifically, the co-signatories note that not all data protection and privacy laws provide for "public interest", research or statistical purposes as an exception to the requirement for consent or as a lawful basis for the processing of personal data. Further, where such exceptions do exist, there may be limitations on the scope of their application.

26. The co-signatories also recognize that, where it is lawful to allow large-scale access or collection, APIs can represent a further safeguard against unlawful scraping. While APIs are not impenetrable, they can afford the host greater control over the data on its platform and facilitate detection and mitigation of unauthorized access, via the use of credentials as well as logging and monitoring of associated activity.

SMC usage of scraped data and data from their own platforms for AI development

27. The co-signatories took the opportunity presented by this initiative to engage with SMCs about their own scraping of data and use of scraped data sets to train their Large Language Models, which present not only opportunities for innovation but also significant privacy risks.

28. Based on what was learned through these engagements, the co-signatories wish to remind SMCs and other organizations who may use scraped personal data or data collected from their own platforms for the development, operation and deployment of generative AI systems, that they must comply with data protection and privacy laws, as well as any other AI-specific laws where they exist. The co-signatories also call on these organizations to comply with privacy and data protection principles

---

Upon a reasoned request from the Digital Services Coordinator of establishment, providers of very large online platforms or of very large online search engines shall, within a reasonable period, as specified in the request, provide access to data to vetted researchers who meet the requirements in paragraph 8 of this Article, for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union, as set out pursuant to Article 34(1), and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 35.

like those detailed in the 2023 Global Privacy Assembly Resolution on Generative Artificial Intelligence Systems and other international guidance[7]. Specifically, the co-signatories note that data protection and privacy laws regulate whether and to what extent the collection and use of personal data for AI development is lawful.

Conclusion

29. Since the release of the initial statement, unlawful data scraping has gained increasing attention, in part due to the rapid emergence and deployment of generative AI systems. Data scraping has also been, and continues to be, widely discussed globally both by data protection authorities and industry.
30. The co-signatories wish to recognize the work of the individual data protection authorities that have produced guidance[8] to address practices related to data scraping. In this guidance, we note the common theme that publicly accessible personal data is generally subject to data protection and privacy laws and should be adequately protected against unlawful scraping.
31. The co-signatories also want to emphasise their expectation that all companies, not just SMCs, protect the publicly accessible personal information that they host against unlawful scraping. Failure to implement adequate safeguards in compliance with applicable laws could result in regulatory intervention, including enforcement action.
32. The co-signatories also wish to remind those engaged in data scraping, as well as SMCs and other organizations who use data from their own platforms to train AI, that they should implement measures to ensure that their data practices comply with data protection and privacy laws.
33. Data scraping is a complex, broad and evolving issue that is, and will stay on the radar of data protection authorities. It should also be a focus for other stakeholders that have a role in protecting privacy, including those with whom we engaged in the course of this initiative. The co-signatories will continue to work to promote compliance in this area, including via future engagement with concerned stakeholders, complementary policy development, public education campaigns, and enforcement[9], including collaborative enforcement.

---

[7] See the Roundtable of G7 Data Protection and Privacy Authorities 2023 Statement on Generative AI, the Hiroshima Process International Code of Conduct for Advanced AI Systems and others.

[8] The Dutch DPA (Autoriteit Persoonsgegevens) issued guidelines and the Italian DPA (Garante Per La Protezione Dei Dati Personali) issued instructions to defend personal data from web scraping. The UK Information Commissioner's Office consultation on generative AI and data protection, including web scraping to train generative AI.

[9] Joint investigations of Clearview AI, Inc. by: the Office of the Privacy Commissioner of Canada, the Commission d'accès à l'information du Québec, the Information and Privacy Commissioner for British Columbia, and the Information Privacy Commissioner of Alberta; and by the UK Information Commissioner's Office and the Office of the Australian Information Commissioner.

34. Meanwhile, the co-signatories encourage SMCs to continue to collaborate with each other and with other stakeholders to share knowledge and strategies and develop solutions to address and respond to this common threat.
35. The co-signatories wish to thank the SMCs and industry stakeholders who demonstrated openness in discussions with regulators. This enabled the co-signatories to develop and share their expectations without the need for formal, resource-intensive enforcement action, to the benefit of all.

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

Carly Kind Privacy 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 INAI (National Institute for Transparency, Access to Information and Personal Data Protection) Mexico
Brent R Homan Commissioner ODPA (Office of the Data Protection Authority) Guernsey	**Mar España Martí** Director AEPD (Agencia Española de Protección de Datos) Spain
Robert Chanas Président CCIN (Commission de Contrôle des Informations Nominatives) Monaco	**Gilad Semama** Commissioner Privacy Protection Authority Israel