Comments of the Alliance for Responsible Data Collection (ARDC)

On EDPB Guidelines 03/2026 on Web Scraping in the Context of Generative AI

I. Introduction

The Alliance for Responsible Data Collection (ARDC) welcomes the opportunity to comment on the European Data Protection Board's Guidelines 03/2026 on web scraping in the context of generative AI. ARDC represents the interests of a diverse group of individuals and organizations that use automation to collect public web data. Common uses of web scraping that pre-date the development of generative AI include the collection of commercial data for price point comparisons and market intelligence, financial analysis, real estate valuations, academic and scientific research, journalism, detecting physical security and cybersecurity threats, government and political accountability, tracking hate speech, and archiving the public internet for historical purposes.

ARDC supports the responsible collection and use of publicly available web data, including appropriate safeguards to protect personal data and the rights of individuals. We recognize the importance of applying the General Data Protection Regulation (GDPR) to the processing of personal data collected through web scraping and support the EDPB's efforts to clarify the responsibilities of organizations engaged in these activities.

At the same time, ARDC is concerned that certain recommendations in the proposed Guidelines extend beyond the protection of personal data and risk establishing broad restrictions on automated access to publicly available information. In particular, the recommendation that data collectors exclude websites based on anti-automation preferences expressed via robots.txt, ai.txt, CAPTCHA, or similar mechanisms fails to distinguish between the website operator's preferences and the rights of individual data subjects under the GDPR. The GDPR should not become a mechanism through which website operators can transform their preferences concerning automated access into regulatory restrictions on the collection of publicly available information, including non-personal data, for purposes unrelated to data protection.

ARDC respectfully urges the EDPB to clarify this distinction and ensure that its final Guidelines remain appropriately focused on the protection of personal data.

II. Areas of Agreement with the Proposed Guidelines

ARDC agrees with several important principles reflected in the proposed Guidelines.

A. Recognition of Web Scraping as an Established Data Collection Technique

ARDC agrees with the Guidelines' characterization of web scraping as a commonly used technique involving automated tools to extract and store information from publicly available web services.

ARDC also welcomes the distinction between targeted and untargeted scraping, including web crawling, in which automated software discovers and follows links to identify available information. Recognizing the different methods and purposes of automated data collection is important to developing an appropriate data collection framework.

B. Defining Collection Criteria and Implementing Appropriate Safeguards

ARDC supports the recommendation that organizations establish their data collection criteria in advance, including identifying the categories of information to be collected and excluded.

We likewise support transparent collection practices, appropriate safeguards for personal data, and processes for extraction, cleansing, structuring, and storage that take account of applicable data protection obligations. Each of these concepts is central to ARDC's Technical Standards and Governance Guidelines.

C. Distinguishing the Responsibilities of Data Collectors and AI Developers

ARDC agrees with the distinction in Section 3.2.1, paragraph 19, between the responsibilities of organizations collecting data and those subsequently using that data to develop generative AI systems. As the EDPB Guidelines (and ARDC guidelines) recognize, a scraper is not, in principle, responsible for an AI developer's subsequent reuse of collected data.

III. Concerns Regarding the Recommendation to Exclude Websites Based on Expressed Anti-Automation Preferences

ARDC's principal concern relates to paragraphs 11, 37, and 64 of the proposed Guidelines, which recommend excluding collection of data from websites that oppose automated collection of public web data through mechanisms such as robots.txt, ai.txt, and CAPTCHA.

A. The Reasonable Expectations of Data Subjects Should Not Be Equated with Website Operators' Preferences.

ARDC respectfully disagrees with treating website operators' anti-automation preferences as a basis for excluding websites under the GDPR's data minimisation framework. The reasonable expectations of individuals who make information publicly available should not be inferred solely from the preferences of the websites hosting that information. Most individuals who publish information on publicly accessible websites do not control the website's robots.txt or ai.txt files, and many may be unaware of their existence or function. New signals, like ai.txt, are not broadly adopted or known even to website operators. Other mechanisms, like CAPTCHA, may be implemented for operational or commercial purposes unrelated to the privacy expectations of individuals who may not know whether a particular website uses CAPTCHA.

Nor does the existence of a robots.txt file that "Disallows" automated access provide reasonable assurance that an individual's publicly posted data will not be scraped and used for generative AI. In fact, some websites' terms of service explicitly advise their users that the website operator may license their publicly available information to AI developers for AI training while at the same time posting a robots.txt file that opposes automated access to all publicly posted content. For example, Reddit's terms of service grant Reddit a license to use user's content "to train AI and machine learning models" (Par. 5), while its robots.txt file purports to disallow all automated scraping ("User-agent: * Disallow: / "). In other words, an assumption that an individual reasonably expects his or her publicly posted content not to be used for generative AI cannot be based on the existence of a robots.txt "Disallow".

Fundamentally, website operators and the individuals whose personal data appears on their websites are distinct parties with potentially different interests. Website operators may oppose automated

collection for commercial, competitive, or operational reasons entirely unrelated to protecting their users' personal data. Their preferences may not reflect the wishes, expectations, or rights of the individuals whose information appears on those websites.

B. Data Minimisation Concerns Personal Data, Not the Method of Collection

Whether information constitutes personal data under the GDPR depends on its relationship to an identified or identifiable individual, not on whether it is accessed manually or through automated tools.

The method of collection does not alter the underlying nature of the information or, by itself, establish that its collection infringes an individual's rights.

Article 5(1)(c) requires that personal data be limited to what is necessary for the purposes of processing. A website operator's anti-automation preference does not determine whether particular information constitutes personal data, whether its processing is necessary, or whether appropriate safeguards have been implemented.

The relevant inquiry should focus on the nature of the personal data, the purposes and circumstances of processing, and the safeguards necessary to protect individuals—not whether a website operator prefers to prohibit automated collection.

C. Blanket Restrictions Disproportionately Affect Targeted Commercial Data Collection

The Guidelines recognize the fundamental differences between untargeted web crawling, large-scale scraping to assemble datasets for generative AI, and targeted collection of specific commercial information. While the Guidelines focus on data scraping for generative AI development and acknowledge that it does "not address the role of data brokers who scrape data and just hold data sets available for other organizations, and whose intention is not to train AI models themselves," the practical impact of the guidelines on commercial scraping cannot be ignored.

A substantial amount of web scraping is directed at publicly available commercial information such as product prices, inventory, customer sentiment, and market trends. Businesses rely on automated collection of this information for price-point comparisons, market intelligence, and competitive analysis.

These activities are especially important to small and medium-sized businesses that depend on automated tools to compete with larger companies in today's global economy. Such targeted collection frequently does not involve processing personal data at all.

Yet website operators' generalized preferences against scraping often fail to distinguish between large-scale AI crawling and targeted collection of non-personal commercial information.

Guidance that endorses website exclusions based on anti-automation preferences risks disproportionately harming smaller businesses, researchers, and other organizations engaged in legitimate activities unrelated to the processing of personal data.

Although the Guidelines "focus" on scraping internet data for generative AI development, as described in Paragraphs 1-6, they should not create a broader regulatory expectation that website operators' preferences determine whether publicly available commercial information may be collected.

V. Recommended Clarifications to the Final Guidelines

ARDC respectfully recommends that the Guidelines explicitly state that they do not apply to scraping internet data that does not include personal data under the GDPR.

ARDC further recommends removal of the sixth bullet of paragraph 37 and the third bullet of paragraph 64 that controllers exclude websites solely because website operators oppose automated access to and collection of their content.

Instead, the Guidelines should focus on whether the processing of personal data is lawful, necessary, proportionate, and subject to appropriate safeguards.

As an alternative to removing the sixth bullet of Paragraph 37, ARDC suggests revising it to read:

- *exclude from the collection personal data that the data subject has posted on a non-public portion of a website, such as behind a password-protected log-in; consider whether the website has advised its users in the websites' terms of service that the users' personal data posted on a non-public portion of the site will not be scraped, licensed, or sold without the users' express consent. Website operators' general preferences regarding automated collection, including those expressed through CAPTCHA, robots.txt or ai.txt files, do not, by themselves, establish restrictions under the GDPR or determine whether particular personal data may lawfully be processed.*

ARDC further recommends clarifying in third bullet in paragraph 64 and Example 5 that website operators' preferences concerning automated collection should not, without additional relevant circumstances such as clear language in the websites' terms of service advising the user that personal data posted on the publicly available portions of the website will not be scraped, sold, or licensed.

As an alternative to the existing third bullet in paragraph 64, ARDC suggests revising it to read:

- *the terms of service available to users of the scraped website, for example, whether the website advises users that personal data users post on non-public portions of the website will or will not be scraped, sold, or licensed for use by third parties, used to improve the website operator's products and services, or used to train generative AI models. Website operators' general preferences regarding automated collection, including those expressed through CAPTCHA, robots.txt or ai.txt files, do not, by themselves, establish restrictions under the GDPR or determine whether particular personal data may lawfully be processed.*

ARDC further alternatively proposes the following revision to Example 5 under Paragraph 64:

> **Example 5:** *If a person uploads their data to a content sharing platform whose terms of service clearly advises users that personal data posted on non-public portions of the website will not be scraped, sold, or licensed, and the website implements measures that prevent any scraping, sale, or use of the users' personal data without their express consent, data subjects can reasonably expect that personal data posted on non-public portions of a website will not be used for generative AI.*

VI. Conclusion

ARDC supports the EDPB's efforts to provide clear guidance on the responsible processing of personal data in generative AI development.

We agree that organizations collecting personal data should define appropriate collection criteria, implement safeguards, minimize unnecessary processing, promote transparency, and respect the rights of individuals.

However, protecting personal data and enforcing website operators' preferences regarding automated access are fundamentally different objectives.

The GDPR should not be used to transform website operators' unilateral preferences into restrictions that extend beyond the protection of personal data, potentially restrict lawful access to information and disadvantage small and medium-sized businesses that depend on automated data collection to compete.

ARDC respectfully urges the EDPB to refine its recommendations to preserve this distinction, protect the rights of data subjects, and avoid unnecessarily restricting responsible access to publicly available information.

Respectfully submitted,

Jo Levy
Chair,
Alliance for Responsible Data Collection