AI Chamber Position Paper for the Public Consultation on EDPB Guidelines 03/2026 on Web Scraping in the Context of Generative AI

AI Chamber submits its comments on the European Data Protection Board's (EDPB) draft Guidelines 03/2026 on web scraping in the context of generative artificial intelligence.

The Guidelines should clearly distinguish between what the GDPR requires and which measures remain voluntary. Their purpose should be to reduce uncertainty without creating additional obligations or leaving scope for such obligations to be inferred from general recommendations.

This is particularly important for small and medium-sized enterprises (SMEs), which do not have extensive legal and technical teams. Businesses should be able to understand and apply the requirements without having to piece together numerous documents or anticipate how individual authorities will interpret non-binding recommendations.

1. Elements that should be retained

AI Chamber supports the distinction between the responsibilities of an entity collecting data under instructions and those of an entity acting independently. The confirmation that a data collector is, in principle, not responsible for the purchaser's subsequent reuse of the data is also important.

We welcome the clarification in paragraph 36 that data minimisation does not preclude training models on large datasets. This principle should not be interpreted as a standalone quantitative restriction on data processing, irrespective of its purpose.

We also support the recognition that special categories of personal data may be collected unintentionally despite safeguards, and the approach encompassing measures at successive stages of model development and use.

Practical examples should be retained and expanded where uncertainty remains. Recording collection dates, using reliable sources and verifying accuracy can be useful practices, provided that they are not presented as additional procedures required in every case.

2. Lawful access to data as a principle guiding the application of the Guidelines

The Guidelines should expressly recognise lawful access to publicly available data and its responsible use as a value to be taken into account when interpreting their recommendations. Data protection should not lead to restrictions beyond the requirements of the GDPR.

The draft recognises that generative AI development requires large volumes of data from diverse sources. The assessment of web scraping should also consider its importance for innovation, competition and the development of services. Restricting web scraping will adversely affect the competitiveness of European businesses and thereby prevent the achievement of important EU objectives, such as strategic autonomy in AI.

The broader applications of this technology include search engines, cybersecurity, research, market intelligence and price comparison services. These illustrate the economic and societal importance of access to information, although the detailed comments in this position paper concern the scope of the Guidelines, namely generative AI development.

Increasing data availability and reuse is a policy objective reflected in Regulation (EU) 2023/2854, Regulation (EU) 2022/868 and Directive (EU) 2019/1024. Mario Draghi's report, "The Future of European Competitiveness", also highlights the importance of innovation and digital transformation for Europe's competitiveness.

Where processing complies with the GDPR, the choice of web scraping as a method of data collection should not itself create an additional regulatory barrier.

3. The Guidelines should not create additional obligations

European businesses operate within a framework comprising legislation, guidance, supervisory interpretations and differing national practices. Further recommendations may increase costs and uncertainty if their status is not clearly defined.

We recommend that each recommendation make clear whether it describes an obligation arising under the GDPR or merely a possible means of fulfilling that obligation. This distinction should be applied consistently throughout the document. Optional measures should not become de facto conditions for demonstrating compliance.

This applies in particular to:

- crawler characteristics and the level of detail required when identifying sources - paragraph 30;
- publication of domain names and URLs in a searchable format - paragraph 31;
- information about the supplier of an acquired dataset and the conditions under which the data was collected - paragraph 32;

- public disclosure of a data protection impact assessment - paragraph 34;
- collection criteria, data mapping and inventories, consideration of synthetic data, and filters - paragraph 37;
- filtering, replacement of data with synthetic data, anonymisation and pseudonymisation - paragraph 38;
- publication of updated lists of websites - paragraph 66(e);
- specific measures to limit memorisation, regurgitation and the vulnerability of models to attacks, including deduplication - paragraph 66(i).

We do not question statutory obligations, including the provision of information about data sources or the performance of an impact assessment where required by law. Our concerns relate to the expansion of those obligations through the Guidelines by prescribing additional information, publication or particular technical methods.

If the EDPB retains individual recommendations, they should remain examples to be adapted to the circumstances. Businesses should be able to use other effective measures that meet the requirements of the GDPR.

Short, voluntary templates, checklists and examples would be helpful for SMEs. They should simplify compliance without creating another layer of documentation. Businesses should also be permitted to use existing documents where these contain the necessary information. Consideration should be given to using IT tools capable of automating compliance processes.

4. Scope of application and actual risk

We support the distinction between personal and non-personal data in paragraphs 12-13, including in relation to mixed datasets. This distinction should be maintained consistently throughout the Guidelines.

Information about products, prices, companies or technical matters may fall outside the scope of the GDPR where it does not relate to an identified or identifiable natural person. Using the same tool to collect different categories of information does not change their nature.

Web scraping should not be treated as inherently high-risk processing. The assessment should consider the type of data, the purpose, the effects on individuals and the safeguards. Large-scale collection of sensitive information requires a different assessment from the use of ordinary, publicly available information in a lower-risk application.

Scale and automation may influence the assessment, but should not, on their own, determine its outcome or trigger additional requirements.

5. Legitimate interests as an accessible legal basis in practice

Paragraph 43 and paragraphs 46–66 recognise and discuss reliance on legitimate interests. The Guidelines should clearly treat legitimate interests as one of the key legal bases that is available in practice for web scraping, provided that the conditions for its application are met.

Clear guidance is needed on assessing the controller's interest, the necessity of processing, including from the perspective of economic necessity, individuals' reasonable expectations, risks and safeguards. This should enable businesses to assess a specific project rather than leave them solely with a catalogue of reservations.

Uncertainty or an excessively restrictive approach may discourage lawful data use and weaken the competitiveness of European companies. Practical examples should also cover small businesses fine-tuning models and using acquired datasets, which is the most common scenario in the European economy.

Pre-collection objection mechanisms, described in paragraph 66(f), may constitute an additional safeguard. Their voluntary nature should be expressly preserved.

6. Clear conditions for the exemption from individual notification

The Guidelines should precisely explain the availability of the exemption under Article 14(5)(b) GDPR where a controller collects data from multiple public sources and has neither contact details nor a reasonable means of reaching the individuals concerned. This is a common and economically important scenario.

Paragraph 27 cautions against routinely relying on the exemption, while Example 2 illustrates its application in AI development. This relationship should be clarified, with explicit confirmation that the exemption is also available in commercial activities where justified by the circumstances and safeguards. The criteria in paragraph 28, including the number of individuals, the age of the data and the protective measures adopted, are helpful in this regard.

Controllers should not be expected to collect additional contact details solely to provide a notice where doing so would involve disproportionate effort and increase privacy risks. Where the purpose of processing does not require the identification of an individual, as is generally the case when data is used for AI, Article 11(1) GDPR, referred to in paragraph 29, should also be taken into account.

7. Transparency without publishing detailed maps of personal data

We maintain our position that no additional disclosure requirements should be introduced. If recommendations on publishing sources are retained, they should expressly take account of proportionality, confidentiality and the risks arising from publication itself.

A public list of URLs of pages containing personal data may facilitate malicious data collection, profiling and re-identification. Publishing such a list may itself constitute processing of personal data. Moreover, a complete list, even in a searchable format, may not best address the needs of an individual seeking information about a particular page.

We therefore propose distinguishing publicly available information from detailed records accessible to the supervisory authority. Public information could include source categories, domains, collection periods, methodology and explanations for omissions. Full URLs and collection records could be made available to the authority upon request.

An alternative to disclosing the entire list should be a tool allowing individuals to check whether a particular website or page was included in data collection. This would provide useful information without publishing a complete map of sources.

Information about the crawler should be limited to what enables individuals to understand the processing and exercise their rights. It should not include operational or technical details that do not serve these purposes.

Similarly, if the recommendation to publish a data protection impact assessment is retained, publication of a summary should be permitted instead of the full document, protecting confidential information and details of security measures.

8. Clear meaning of objection signals

Paragraphs 37 and 64 refer to mechanisms such as robots.txt, ai.txt, CAPTCHA and authentication. The Guidelines should clearly specify whether disregarding a particular signal constitutes a GDPR infringement in itself or is merely one element of a broader assessment.

These mechanisms differ in purpose and scope. Directives in robots.txt may apply to a particular crawler, while CAPTCHA or authentication may protect a specific part of a website. They should not automatically be interpreted as opposition to every method of collecting all content within a domain.

If the EDPB retains these recommendations, it should distinguish explicit reservations and access controls applying to specific content from general technical instructions. Non-standardised files, such as ai.txt, should not have independent regulatory significance without sufficient standardisation. These recommendations must not create additional obligations that do not arise under the GDPR.

AI Chamber remains ready to contribute to Guidelines that protect individuals and provide businesses with clear and predictable rules. The final document should reduce divergent interpretations and enable compliance without creating new requirements, particularly for SMEs.