AmCham EU
SPEAKING FOR AMERICAN BUSINESS IN EUROPE

Consultation response

EDPB public consultation on draft Guidelines 03/2026 on web scraping in the context of generative AI

AmCham EU speaks for American companies committed to Europe on trade, investment and competitiveness issues. It aims to ensure a growth-orientated business and investment climate in Europe. AmCham EU facilitates the resolution of transatlantic issues that impact business and plays a role in creating better understanding of EU and US positions on business matters. Aggregate US investment in Europe totalled more than €4 trillion in 2023, directly supports more than 4.6 million jobs in Europe, and generates billions of euros annually in income, trade and research and development.

American Chamber of Commerce to the European Union
Speaking for American business in Europe

Avenue des Arts/Kunstlaan 56, 1000 Brussels, Belgium • **T** +32 2 513 68 92
info@amchameu.eu • amchameu.eu • European Transparency Register: 5265780509-97

Executive summary

Clear, proportionate and technically feasible rules are critical to foster European competitiveness, innovation and the successful development and deployment of AI. The European Data Protection Board's (EDPB) draft Guidelines 03/2026 on web scraping in the context of generative AI can support this legal certainty by ensuring a harmonised application of the General Data Protection Regulation (GDPR) across the European Economic Area. However, certain provisions of the current draft Guidelines require adjustments to ensure alignment with the technical realities of AI development. These adjustments include:

- Ensuring that the disproportionate effort condition applies whenever the objective conditions are met, avoiding technically unfeasible requirements for detailed data source documentation and maintaining the confidentiality of Data Protection Impact Assessments (DPIAs).
- Clarifying the application of legitimate interest by acknowledging the necessity of broad, untargeted web crawling for the development of general-purpose and foundational AI models, removing technically impracticable global opt-out lists and ensuring that safeguards are assessed proportionately.
- Aligning data accuracy and minimisation requirements with technical realities by avoiding impractical verification and documentation requirements, recognising the importance of diverse training datasets and guaranteeing consistency with data requirements under the AI Act.
- Adopting a risk-based approach for special category data (SCD) by replacing absolute deletion requirements with a 'reasonable steps' standard and recognising that safeguards during data collection and deployment-phase filtering can jointly mitigate risks associated with SCD.
- Clarifying the scope of the Guidelines by distinguishing large-scale scraping for AI training from user-initiated retrieval and security/integrity-related crawling and ensuring that GDPR obligations apply only when information constitutes personal data.

These changes would complement existing positive clarifications present in the draft Guidelines, including the recognition of legitimate interest as a valid lawful basis for the collection of data for AI model training and the acknowledgement that data mining does not prohibit large-scale data processing. The pragmatic approach to the incidental collection of SCD and to exceptions to transparency obligations in cases of disproportionate effort is also appreciated.

This pragmatic, risk-based approach to data protection and AI development will help achieve the Digital Omnibus's objectives of simplification and efficiency, while supporting European competitiveness.

Introduction

The development of generative AI relies on access to large and diverse datasets, including information collected from publicly accessible online sources. In this context, the EDPB has published its draft Guidelines 03/2026 to clarify how the GDPR applies to web scraping for generative AI.

Harmonised regulatory guidance is essential to provide legal certainty and enable the responsible development of AI technologies across the European Economic Area. To foster European innovation and the successful adoption of AI, it is critical that these rules remain clear and proportionate. While

the draft Guidelines establish positive foundational principles, their recommendations require adjustments to align with the technical realities of AI development.

This consultation response outlines the positive elements of the draft Guidelines, identifies areas where further clarification or adjustments are needed and sets out additional considerations to ensure that the final Guidelines are technically feasible.

Areas of support

The EDPB adopted several pragmatic approaches in the draft Guidelines, which provide necessary legal certainty for AI developers:

- **Legitimate interest and minimisation:** the recognition of legitimate interest (Article 6(1)(f) GDPR) as a valid lawful basis provides greater clarity for the collection of data for AI model training. Similarly, the clarification in Paragraph 36 that data minimisation does not, in itself, prohibit training models on large volumes of data is welcome.
- **Special category data and the GC case:** the pragmatic application of the GC and Others (C-136/17) case law to the incidental collection of SCD is positive, assessing processing within the bounds of a controller's actual 'responsibilities, powers, and capabilities'.
- **Transparency and Article 11:** Paragraphs 25-28 are right to acknowledge that the disproportionate effort exception (Article 14(5)(b) GDPR) can apply to large-scale web scraping. Furthermore, we strongly welcome the confirmation in Paragraph 29 that Article 11 GDPR shields controllers from being required to acquire additional identifying information solely to comply with data subject requests for unstructured scraped data.

Areas of improvement

A. Transparency obligations and the disproportionate effort exception

- **Problem:** While the Guidelines recognise the Article 14(5)(b) exception, the assertion that it should not be 'routinely' relied upon appears unduly restrictive. The application of this exception should depend on the objective facts of each case (eg the vast scale of processing), regardless of frequency. Furthermore, Example #2 suggests excluding data from web pages where subjects are directly identifiable; it is unclear how such pages can be systematically detected at scale. Additionally, Paragraph 31 suggests controllers include granular domain names, searchable URLs and precise dates of collection. For large-scale AI systems, maintaining such exhaustive lists is technically unfeasible. Finally, Paragraph 34 proposes making Data Protection Impact Assessments (DPIAs) publicly available, which poses severe risks to commercial confidentiality and system security.
- **Recommendation:** The EDPB should clarify that the Article 14(5)(b) exception applies whenever the factual conditions are met, removing the restrictive caveat that it cannot be used 'routinely'. Paragraph 31 should be revised to recognise that detailing sources at the category or domain-type level is sufficient. Finally, the EDPB should remove the suggestion to publish DPIAs, clarifying that they remain available to Data Protection Authorities upon request.

B. The legitimate interest balancing test and 'untargeted' crawling

- **Problem:** The balancing test sets an unreasonably high bar. Crucially, the Guidelines imply a preference for 'targeted scraping' with restrictive criteria over 'untargeted internet crawling'.

This fundamentally misunderstands how general-purpose and foundational AI models are developed; they require broad, diverse, untargeted data ingestion to learn human language and context. Furthermore, Paragraph 66(f) suggests creating global opt-out lists. Centralised global opt-out lists are technically unworkable and place an impossible identification burden on controllers, as they would require data subjects to provide information identifying them across specific websites or online platforms. This could itself entail the collection and centralisation of additional identifying information, thereby creating further privacy risks for the very individuals the mechanism is intended to protect. Paragraph 65 further states that mitigating measures should not be confused with measures already legally required under the GDPR. This is overly rigid: measures that materially reduce risks to data subjects should be capable of carrying weight in the balancing test even where they also implement existing GDPR obligations. Relatedly, machine-readable signals such as robots.txt or ai.txt may be relevant to reasonable expectations, but they should not be treated as determinative GDPR compliance mechanisms or absolute bars to collection. Finally, Example 7 integrates compliance with the Copyright Directive into the GDPR balancing test, creating regulatory incoherence.

- **Recommendation:** The EDPB should explicitly acknowledge that broad, untargeted web crawling is a necessary and proportionate technical reality for the training of general-purpose and foundational AI models, and should not be penalised in the necessity threshold. The expectation of global opt-out lists should be removed, clarifying that machine-readable signals (eg robots.txt or ai.txt) are relevant factors in the balancing test rather than determinative GDPR opt-outs. The Guidelines should also clarify that safeguards which materially mitigate risks may be considered in the balancing test even where they implement existing GDPR obligations, and should provide clearer practical examples of when proportionate safeguards satisfy the test. Example 7 should be amended to separate copyright law from data protection assessments.

C. Accuracy, data minimisation, and 'reliable sources'

- **Problem:** Under the accuracy principle, the Guidelines recommend scraping from 'reliable sources' (eg official registries) and suggest controllers timestamp data to track its age and conduct factual verification audits on random samples prior to training. This imposes traditional database expectations onto the training of large-scale AI systems, which is technically impractical. For datasets of this scale, random spot-checking cannot provide meaningful confidence in the accuracy of the overall corpus and, in some cases, factual verification may itself require collecting additional information about individuals solely for verification purposes. Artificially limiting scraping to 'official' sources risks creating heavily biased models. Regarding minimisation, Paragraph 37 recommends a 'data mapping and inventory process', creating a duplicative burden already covered by Records of Processing Activities (ROPA). Furthermore, limiting data collection in line with the Guidelines may make it more difficult to meet the data quality requirements under Article 10 of the AI Act, which require training datasets to be sufficiently diverse, representative and complete.
- **Recommendation:** The EDPB should remove expectations for granular timestamping and pre-training factual verification audits, which are incompatible with large-scale generative AI ingestion. The EDPB should broaden 'reliable sources' to acknowledge the necessity of diverse internet data. The Guidelines should also state that existing ROPA documentation satisfies the data mapping recommendations. To ensure coherence across the EU legal framework, the EDPB should interpret the GDPR's data minimisation principle in a manner consistent with the

data quality requirements under Article 10 of the AI Act. This would help minimise the risk of conflicting requirements under the two legislative frameworks.

D. Feasibility of safeguards for special category data (SCD)

- **Problem:** While the *GC and Others* framework is positive, the specific conditions in Paragraph 74 are absolute and rigid, omitting the crucial concept of 'reasonable steps'. Paragraph 74(ii) expects controllers to ensure incidentally collected SCD is 'deleted from datasets immediately'. Pre-scrubbing the entire internet to guarantee the absolute absence of SCD presents disproportionate technical challenges given current state-of-the-art technologies. Furthermore, the Guidelines create an unworkable double compliance burden: expecting controllers to perfectly exclude SCD at the point of collection, while also requiring state-of-the-art content filters during live deployment to prevent the model from revealing SCD.
- **Recommendation:** The EDPB should add the concept of 'reasonable steps' into Paragraph 74, reframing it from an absolute guarantee of deletion to a risk-mitigation standard. The Guidelines should also recognise that a combination of ingestion-phase safeguards and deployment-phase filtering constitutes sufficient mitigation, rather than demanding perfection at both stages independently.

E. Scope and applicability of the Guidelines

- **Problem:** The Guidelines do not clearly distinguish the GDPR threshold question and different forms of automated retrieval. First, web content should not be treated as personal data by default merely because it may contain a number, image or other reference to an individual; Recital 26 requires a contextual assessment of identifiability from the controller's perspective. Second, user-directed, inference-time retrieval (including agentic browsing, tool use and retrieval-augmented generation) differs materially from large-scale pre-training scraping: it is typically targeted, user-initiated and used transiently to respond to a specific request. Crawling for integrity and security purposes likewise serves a distinct purpose, such as detecting malicious content, fraud or unsafe links. Applying training-specific requirements indiscriminately to these activities risks creating legal and technical uncertainty. Where data collected for integrity or security purposes is technically and organisationally ring-fenced from use for GenAI training, that separation demonstrates purpose limitation in practice.
- **Recommendation:** The EDPB should clarify that the Guidelines apply only where the information at issue constitutes personal data under the GDPR's contextual identifiability standard. It should also clarify that real-time, user-initiated retrieval for immediate response generation falls outside the training-time scraping framework, or at minimum is subject to a distinct and proportionate analysis, and that integrity/security crawling should be assessed according to its separate purpose. Technical and organisational barriers preventing reuse of security-crawled data for GenAI training should be recognised as relevant safeguards. Any subsequent reuse of data obtained through user-directed retrieval for model improvement should be assessed separately according to the nature and scale of that secondary processing.

Additional observations

As the European Union negotiates the Data and Privacy Digital Omnibus, it is vital that regulatory bodies deliver a cohesive, predictable legal environment for businesses operating in Europe. The co-legislators are currently working to establish horizontal, harmonised rules on the processing of data for AI development, including the handling of incidental data and the parameters of legitimate interest.

The EDPB should therefore ensure that Guidelines 03/2026 do not pre-empt, complicate or contradict the ongoing democratic negotiations surrounding the Digital Omnibus. The final Guidelines must remain sufficiently flexible and risk-based to align with the ultimate political agreement and the overarching goal of the Digital Omnibus: to foster a competitive, innovative European AI ecosystem underpinned by legal certainty.

Conclusion

AmCham EU remains committed to working with the EDPB to ensure the highest standards of data protection while enabling the European AI ecosystem to thrive. We urge the EDPB to adopt a pragmatic, risk-based approach in the final version of these Guidelines. Our members stand ready to provide further technical expertise to assist the Board in finalising workable rules.