Ecommerce Europe feedback on EDPB guidelines 03/2026 on web scraping in the context of generative AI

Ecommerce Europe is the united voice of the European Digital Commerce sector, representing the interests of companies selling goods and services online to consumers in Europe. We welcome the European Data Protection Board's (EDPB) efforts to ensure coherent clarifications on the legal and technical implications of web scraping for generative artificial intelligence (AI) development under the GDPR. Online retailers are affected in two ways: as operators of websites and platforms whose content may be scraped, and increasingly as buyers or users of scraped datasets and AI tools built on them. Kindly find below our observations on targeted aspects of the guidelines.

Points we welcome

We welcome in particular the confirmation that legitimate interest can be relied on for scraping in the context of generative AI, the confirmation that controllers need not collect additional information solely to identify data subjects in order to handle rights requests, and the clarification that the disproportionate effort assessment under Article 14(5)(b) GDPR is made on the dataset as a whole. We also welcome Example 4, which recognises that where a platform tells its users that content may be scraped, those users can reasonably expect it.

Insufficient guidance on acquired datasets

The draft guidelines discuss situations where an organization obtains a previously scraped dataset from a data broker (para 3), confirm that the scraper and the AI developer are in principle responsible for their own separate processing (para 19), and require the developer to consider the datasets it intends to use (para 21). However, they provide limited practical guidance regarding what due diligence is required, what evidence the company must obtain, when a purchaser may rely on the original collector's assessment, and whether independent verification is always required.

For many of our members licensing datasets from verified sources is a more practical option rather than scraping directly. The Guidelines should therefore provide a due diligence framework for purchasers of scraped datasets, including:

a. minimum documentation requirements;
b. acceptable evidence of lawful collection;
c. allocation of accountability between scraper, broker and purchaser, and;
d. circumstances under which reliance on vendor assessments is sufficient.

Transparency requirements may be unrealistic

The draft acknowledges that Article 14 notices may be impossible or involve disproportionate effort but then appears to expect publication of extensive information including categories of data, sources, crawler characteristics and collection methodology. For organizations buying datasets rather than scraping themselves, some of this information may not be available due to intellectual property restrictions, trade secrets, or vendor confidentiality obligations.

The Guidelines should clarify the level of granularity required in public notices (paras 30-32) where data were acquired from a third party, source inventories are commercially confidential, or disclosure would reveal security-sensitive crawler configurations. For acquired datasets, para 32 already recommends linking to the website of the original scraper. The Guidelines should confirm that a link to the original scraper's notice, together with a short explanation of how the data was obtained, satisfies the source information requirement for the purchaser. Further examples of acceptable transparency practices would improve legal certainty.

Undefined meaning of "reliable sources"

The Guidelines recommend scraping from "reliable sources" and validating data before model training (para 42). The examples given (a government registry or a verified company file) suit factual records about individuals but say little about the wider range of content used to train general models. This matters because "reliable source" could become a de facto compliance obligation, even though para 42 frames it as a recommendation to be followed "to the greatest extent possible"/

The Guidelines should confirm that scraping from "reliable sources is good practice rather than a condition of lawfulness" and explain whether reliability relates to data accuracy, authenticity of publication, data protection compliance, source governance controls, or a combination of these factors.

Inadequate treatment of anonymisation thresholds

The Guidelines repeatedly recommend anonymisation, pseudonymisation, or replacement with synthetic data. However, they do not provide meaningful guidance on when anonymisation is considered effective, how to assess re-identification risks in foundation model training, and whether post-training model leakage risks affect anonymisation status.

The Guidelines should provide practical criteria for assessing anonymisation effectiveness in AI training contexts, cross-referring to the EDPB's draft Guidelines 02/2026 on anonymization so that a single standard applies and clarifies whether residual memorisation risks influence the assessment of anonymisation under GDPR.

Lack of clarity regarding incidental collection of special category data

The EDPB relies on the GC & Others (C-136/17) judgment and suggests that incidental, residual collection of special category data may in some circumstances be acceptable if safeguards exist (paras 72-74). Footnote 44 describes incidental and residual collection as data the controller has no intention of collecting but which is collected despite preventive measures. However, the draft does not explain how a controller

shows that its collection meets that description, or what preventive measures are sufficient. This could become a major point because web-scale datasets often inevitably contain health, political, religious, or other Article 9 data.

The Guidelines should clarify how a controller can demonstrate that collection is incidental rather than systematic and give examples of combinations of technical and organizational measures that would generally satisfy the expectations in para 74.

Accountability expectations appear stronger than the practical ability to identify data subjects

One recurring tension in the draft is that controllers remain accountable for GDPR compliance, yet the EDPB simultaneously recognizes that controllers may not know which personal data or which individuals are included in large-scale datasets (para 14). This creates a practical tension, which para 29 partly addresses by confirming that controllers need not collect additional information solely to identify data subjects.

For example: How are rights requests handled? How should retention schedules be applied? How can accuracy obligations be applied?

The Guidelines should provide sector-specific accountability examples demonstrating compliance where controllers cannot identify individual data subjects, datasets contain billions of records, or rights requests cannot be linked to specific records without collecting additional personal information.

No risk-based distinction between targeted and untargeted scraping

The EDPB introduces the distinction between targeted and untargeted scraping (para 8) and links breadth of collection to the necessity test (para 52) and to the level of interference (para 58) but does not explain how targeted collection should be reflected in the assessment.

For organizations, the natural expectation would be targeted scraping means lower risk and untargeted scraping a higher risk, but the draft implies this but never explicitly states the consequences.

The Guidelines should confirm that narrowly targeted collection is a relevant factor in favour of the controller in the necessity and balancing assessments and give an example of how it is weighed.

Online retailers both operate websites, marketplaces and review platforms whose content, including customer reviews and seller profiles, may be scraped and routinely collect publicly available data themselves, for example competitor pricing, product reviews and product descriptions. Our members therefore have a direct interest in rules that are clear on both sides.

We support the expectation that controllers exclude from collection websites that clearly oppose scraping those measures such as robots.txt, ai.txt or CAPTCHA, and the recognition in Example 5 that users of such a platform cannot reasonably be expect their data to be scraped for AI development.

Para 64 lists the ‘restrictions imposed by the scraped website’ as a factor in reasonable expectations but refers only to technical signals. In practice, many website operators also set out their position in their terms of service.

The Guidelines should confirm the platform terms of service weight in the data minimization and reasonable expectations assessments where they are (a) express and prominently accessible, (b) specific to AI training or model development, and (c) accompanied by at least one technical measure such as robots/tt, ai-txt or a login requirement. This combined standard avoids treating generic boilerplate as a barrier while giving meaningful effect to a website operator’s deliberate, published choice. It also preserves the ability to collect publicly accessible data for other legitimate purposes, such as competitive monitoring, where none of these conditions are met.

Organizations’ own data

Para 5 states that the Guidelines do not address the processing of an organization’s own personal data. Many retailers fine-tune AI tools, such as customer service assistants, on their own product and customer data rather than on scraped datal. The Guidelines should confirm that such first-party processing is assessed under the general GDPR framework and not under the scraping-specific expectations set out in these Guidelines.