Comments on EDPB Guidelines 03/2026 on web scraping in the context of generative AI

Public consultation, 8 July to 30 October 2026

Submitted by:

Dr. Juan Emmanuel Delva Benavides
Dr. Jorge Antonio Leos Navarro
Dr. Alejandro Paul García Hernández

Research Group (Cuerpo Académico) UDEG-CA-1236 "Law and Technology", Department of Social and Legal Sciences, University Centre for Economic and Administrative Sciences (CUCEA), University of Guadalajara, Mexico

Contributor type: academic / research institution · Country: Mexico

This contribution reflects the academic views of its authors as members of the research group. It does not represent an institutional position of the University of Guadalajara.

1. Introduction

In September 2024 the Dutch Data Protection Authority (Autoriteit Persoonsgegevens) fined Clearview AI €30.5 million for building a database of more than 50 billion facial images scraped from the open web without informing the persons concerned or giving them a way to access their data. Three months later, in December 2024, the Italian Garante per la protezione dei dati personali fined OpenAI €15 million, among other grounds, for processing personal data to train ChatGPT without an adequate legal basis and without sufficient transparency. Both decisions share a premise that the European Data Protection Board (EDPB) now places at the centre of its Guidelines 03/2026: personal data do not lose their protection because they are publicly accessible, and harvesting them at scale is a processing operation that must find its own justification under the General Data Protection Regulation (GDPR).

We welcome the Guidelines and, in particular, the decision to treat web scraping as a processing activity in its own right, distinct from the later phases of model development and deployment already addressed in Opinion 28/2024. The document is right to recall that the accountability principle (Article 5(2) GDPR) applies in full even where compliance is technically demanding (¶14), and that the absence of a robots.txt file does not amount to consent (¶45).

We write from outside the Union, and this vantage point is the reason for most of our comments. The Spanish- and Portuguese-language web is a significant source of training data for generative AI models offered in the European market, and the persons whose data are harvested from it frequently live in Latin America. At the same time, regulators in the region read EDPB guidance closely. Mexico illustrates the point: on 20 March 2025 a new Federal Law on the Protection of Personal Data Held by Private Parties (LFPDPPP) was published, and the supervisory functions of the former National Institute for Transparency, Access to Information and Personal Data Protection (INAI) were transferred to the Secretariat of Anticorruption and Good Government. In a moment of institutional redesign, the criteria

that the EDPB adopts on scraping are likely to travel well beyond the Union. Our comments therefore seek greater legal certainty on points that matter especially for non-EU data subjects and actors.

2. Summary of recommendations

- Scope (¶6): clarify how the Guidelines apply to scraping carried out for scientific research, including by public universities and research consortia, and how the safeguards of Article 89 GDPR interact with transparency and legitimate interest.
- Territorial dimension: add a short section on Article 3 GDPR and Recital 14, covering non-EU developers whose processing falls under Article 3(2) and EU controllers that harvest data concerning persons outside the Union.
- Responsibility along the data supply chain (¶16–20): complement ¶19 with documentation and due-diligence expectations, so that the separation of responsibilities does not turn into an accountability gap.
- Opt-out signals (¶37, ¶45, ¶64, ¶66, Example 7): state that disregarding machine-readable reservations should normally preclude reliance on Article 6(1)(f), and address reservations expressed after the first collection.
- Locale-aware minimisation (¶37–38): require filters calibrated to the languages and identifier formats actually present in the corpus, not only to European formats.
- Irreversibility (¶60): give express weight in the balancing test to the difficulty of erasing personal data from trained models.
- Special categories (¶67–77): narrow the analogy with GC and Others (C-136/17) and clarify the limited role of Article 9(2)(e) GDPR.
- Interplay with the Artificial Intelligence Act (Regulation (EU) 2024/1689): add cross-references to Articles 2(7), 10 and 53(1)(c)–(d).

3. Detailed comments

3.1 Scope: private entities, public bodies and scientific research (¶4–6)

Paragraph 6 limits the Guidelines to scraping by private entities, on the ground that generative AI development is mainly a private-sector activity. The choice is understandable, but it leaves without guidance an actor that is far from marginal: universities, public research organisations and mixed consortia that harvest web data to build research corpora or to train and evaluate models. Paragraph 27 already indicates that the exception of Article 14(5)(b) is particularly relevant to archiving, scientific or historical research and statistical purposes subject to Article 89(1). It bears noting, however, that the Guidelines do not explain when a scraping project qualifies as scientific research, which Article 89 safeguards are expected (pseudonymisation, access controls, purpose limitation of the corpus), or what happens when a research corpus is later licensed to, or re-used by, a commercial developer. The last scenario is common in practice, and it is precisely where the separation of responsibilities set out in ¶19 becomes most fragile.

We suggest replacing the exclusion in ¶6 with a statement that the Guidelines apply mutatis mutandis to public bodies and research institutions, without prejudice to Article 6(1)(e) and Article 89 GDPR, bearing

in mind that the last subparagraph of Article 6(1) excludes Article 6(1)(f) for public authorities in the performance of their tasks. An example of a research corpus subsequently licensed to a private developer would make the point concrete.

3.2 Territorial scope and non-EU data subjects

The Guidelines do not address Article 3 GDPR, and two recurrent situations call for it. The first is a developer established outside the Union that harvests websites to train a model offered in the Union, collecting along the way data of persons who are in the Union, where the processing relates to the offering of goods or services to them or to the monitoring of their behaviour (Article 3(2)). The second is a controller established in the Union that harvests websites whose content mostly concerns persons located elsewhere, for instance in Latin America. In the second case the GDPR applies irrespective of the nationality or place of residence of the data subjects (Recital 14), and the same transparency, minimisation and balancing obligations follow. In our experience, developers do not always draw this consequence and tend to assume that data concerning non-Europeans fall outside the Regulation.

A short paragraph recalling Articles 3(1) and 3(2) and Recital 14, with an example of an EU-established developer harvesting Spanish-language sources about persons in third countries, would close that gap. It would also reinforce international cooperation. The joint statements on data scraping issued in August 2023 and October 2024 by authorities that are members of the Global Privacy Assembly's International Enforcement Cooperation Working Group, signed among others by the then Mexican INAI and, in 2024, by the Spanish Agencia Española de Protección de Datos (AEPD), already express a shared understanding that publicly accessible personal data remain protected.

3.3 Responsibility along the data supply chain (¶16–20)

We agree with the case-by-case approach to roles and with ¶18, under which a scraper acting on the documented instructions of a developer may be a processor. Paragraph 19, however, states that the scraper is not, in principle, responsible for the re-use of the data, and that scraper and developer answer for their own separate processing. Read in isolation, that statement invites a fragmentation of the supply chain. Who answers for the lawfulness of a dataset that was harvested by one company, cleaned and packaged by a second and used for training by a third? Must the developer verify how the data were collected, or may it rely on contractual assurances? And who informs the data subject when none of the three has a direct relationship with that person? Opinion 28/2024 (¶129) indicates that supervisory authorities should take into account whether a controller deploying a model developed by another controller conducted an appropriate assessment to ascertain that the model was not developed through unlawful processing, and ¶21 of the Guidelines already requires the developer to consider the datasets it intends to use. The same logic should be made explicit for third-party datasets.

To that end, ¶19 could add that a scraper making a dataset available to others should document the sources, the collection period, the characteristics of its crawler (the automated program that visits and copies web pages), the filters applied and the opt-out signals respected, in the form of a datasheet; and that a developer re-using a third-party dataset should carry out and document a due-diligence review of its lawful constitution before training, proportionate to the risks. This documentation would also feed the transparency measures of ¶30–31.

3.4 Transparency and the Article 14(5)(b) exception (¶24–34)

We support ¶27, which warns against routine reliance on Article 14(5)(b) outside the purposes of Article 89, and ¶30, which makes public information a measure that the controller must always take, including the precise indication of sources and the characteristics of the crawler. Building on the good practices of ¶31 (domain names and URLs in searchable format, period of collection) and on ¶66(e) (an updated list of scraped websites), we suggest elevating these elements to a minimum, machine-readable content of the public notice, adding the crawler's user-agent string (the identifier a crawler declares to each website) and its published IP ranges, the opt-out mechanisms honoured and a contact point for the exercise of rights. A stable and versioned notice allows data subjects, website operators and authorities to verify the controller's claims over time. Consistency with the public summary of training content that providers of general-purpose AI models must publish under Article 53(1)(d) of the AI Act would allow a single disclosure to serve both regimes.

3.5 Opt-out signals and reasonable expectations (¶37, ¶45, ¶64, ¶66)

The Guidelines rightly state that controllers should exclude websites that clearly oppose scraping through technical measures such as robots.txt, ai.txt or CAPTCHA (¶37), and that such signals reduce the data subjects' expectation of being scraped (¶64). We suggest going one step further: where a website or a data subject has expressed a machine-readable reservation against scraping or AI training, disregarding it should normally preclude reliance on Article 6(1)(f) for the data concerned, rather than counting as one more element of the balancing. Example 7 already refers to the reservation of rights for text and data mining (TDM) under Article 4(3) of Directive (EU) 2019/790; the consequence should be drawn in ¶64. This would align data protection practice with the obligation of providers of general-purpose AI models to adopt a policy to identify and comply with such reservations (Article 53(1)(c) AI Act), without conflating the two regimes.

Two practical questions also deserve an answer. The first concerns reservations expressed after the initial collection: controllers should check them at each re-crawl and before each training run, and exclude from future datasets the data from sources that have since opted out. The second concerns standardisation. The ecosystem of preference signals is evolving, including the draft vocabulary of the AI Preferences (aipref) Working Group of the Internet Engineering Task Force (IETF), and the Guidelines could state that controllers are expected to honour widely used and publicly documented signals, not only robots.txt.

3.6 Locale-aware minimisation and accuracy (¶35–42)

The measures listed in ¶37–38 are well chosen: precise collection criteria, exclusion of sites mainly used by minors and syntax-based filtering of identifiers through regular expressions (patterns that recognise, for example, the structure of an identity number). Their effectiveness, however, depends on the languages and jurisdictions present in the corpus. A filter designed around European identifier formats will not detect the Mexican Unique Population Registry Code (CURP) or the Federal Taxpayer Registry code (RFC), nor Colombian or Chilean national identity numbers. Nor will it detect the conventions of Latin American judicial bulletins, in which the names of parties, including at times minors involved in family proceedings, may appear despite the applicable protections. Keyword filters for special categories of data built for English or a few EU languages perform poorly on regional varieties of Spanish and Portuguese.

We therefore suggest adding to ¶38 that filtering mechanisms should be calibrated and tested against the languages, scripts and identifier formats actually present in the corpus, with documented recall on representative samples, and listing public registers and judicial publications of third countries among the sources that may require exclusion or specific treatment under ¶66(b)–(c).

3.7 Legitimate interest: necessity and balancing (¶43–66)

We support the three-step test and the requirement to identify precisely whether the interest pursued is commercial, public or research-oriented (¶48–50). Three adjustments would strengthen it.

Paragraph 60 acknowledges that, given the current state of the art, personal data cannot easily be deleted from a trained model. This irreversibility should weigh expressly in the balancing: the harder it is to give effect to erasure and objection after training, the stronger the ex ante measures (exclusion, filtering, pseudonymisation) required for the balance to favour the controller. The Guidelines could also set expectations on how objections and erasure requests are carried over to future training runs and to models derived from the original dataset.

Notwithstanding the value of the discretionary and prior right to object, which we welcome among the measures of ¶66(f) together with the opt-out list that may be used before collection, its effectiveness depends on operational conditions: a request channel available in the languages of the data subjects concerned, reasonable response times, confirmation to the requester and a documented process to exclude the data from subsequent datasets. Describing these minimum features would prevent opt-out mechanisms that exist only on paper.

Finally, ¶37 recommends excluding sites mainly used by minors. In view of the specific protection afforded to children (Recital 38) and of the risks of deep-fakes in pictures, videos and speech noted in ¶54 and illustrated for voice data by Example 6, the exclusion of content concerning minors and other vulnerable persons should operate as a strong presumption in the balancing test and not only as a minimisation measure.

3.8 Special categories of personal data (¶67–77)

We agree that Article 9(2) must be met in addition to Article 6, that unjustified residual collection is unlawful (¶69) and that the judgment of the Court of Justice of the European Union (CJEU) in GC and Others (C-136/17) should not be read as a general exemption (¶72). The analogy should nonetheless be narrowed further. The Court's reasoning on the responsibilities, powers and capabilities of a search engine operator rests on a verification carried out ex post, upon request by the data subject, with respect to a specific referencing. Model training differs on two material points: the data are incorporated into the parameters of the model rather than indexed, and, as ¶60 itself acknowledges, they cannot easily be removed afterwards. The control mechanism on which the Court relied is therefore largely unavailable once training has taken place.

Hence, ¶74–75 should state expressly that the ex ante measures of ¶74(i), and their documented effectiveness, are a precondition for relying on the reasoning of C-136/17, and that the measures of ¶74(iii)–(iv) (output filtering, monitoring, unlearning) complement but cannot replace them. It would also be useful to state that the mere public accessibility of information does not mean that the data subject has manifestly made public special-category data within the meaning of Article 9(2)(e), in line with Meta Platforms and Others (C-252/21, paragraph 77), where the Court required that the data

subject intended, explicitly and by a clear affirmative action, to make the data accessible to the general public. For the same reason, the wording of Example 7 ("manifestly made the content public"), used in the Article 6(1)(f) analysis, should be adjusted to avoid confusion with Article 9(2)(e).

3.9 Interplay with the AI Act

The Guidelines do not refer to the AI Act. The obligations of providers of general-purpose AI models concerning training data apply since 2 August 2025, and data governance requirements will apply to high-risk systems. A short section would add legal certainty by recalling that the AI Act shall not affect the GDPR (Article 2(7), save for the specific provisions it lays down on the processing of special categories of data for bias detection); that compliance with the copyright policy and training-content summary of Article 53(1)(c)–(d) does not by itself provide a legal basis under the GDPR, although it can support transparency and accountability; and that data governance under Article 10 for high-risk systems and the minimisation and accuracy measures of the Guidelines should be designed coherently.

4. Drafting suggestions

Paragraph	Issue	Suggested change
¶6	Exclusion of public bodies and research	Apply the Guidelines mutatis mutandis to public and research scraping; add an example of a research corpus later licensed to a private developer.
New ¶ after ¶15	Territorial scope	Recall Article 3 and Recital 14 GDPR; example of an EU controller harvesting third-country sources.
¶19	Risk of an accountability gap	Add documentation duties for dataset providers and due-diligence duties for re-users before training.
¶31	Content of public notices	List minimum machine-readable elements (user-agent, IP ranges, period, sources, opt-outs honoured, contact point); align with the summary under Article 53(1)(d) AI Act.
¶37 / ¶64	Weight of opt-out signals	Disregarding machine-readable reservations should normally preclude Article 6(1)(f); check reservations at each re-crawl and training run.
¶38	Filters calibrated to European formats	Require locale-aware filters tested against the languages and identifier formats in the corpus.
¶60 / ¶65	Irreversibility	Make the difficulty of erasure an express factor requiring stronger ex ante measures.
¶66(f)	Right to object	Describe the minimum operational features of the discretionary and prior right to object.
¶74–75	Analogy with C-136/17	Condition reliance on effective ex ante measures; clarify that public accessibility does not equate to 'manifestly made public' (Article 9(2)(e)); adjust the wording of Example 7.
New section	AI Act	Cross-reference Articles 2(7), 10 and 53(1)(c)–(d) of Regulation (EU) 2024/1689.

5. Concluding remarks

The Guidelines provide a sound framework at a moment when the harvesting of web data for training has become the raw material of generative AI. The Clearview and OpenAI decisions show that European authorities already sanction the most visible abuses; what remains open is the ordinary, daily practice of a supply chain that is global by design, in which European controllers process data of persons outside the Union and non-European developers serve the European market. Viewed in this light, the value of the Guidelines will depend less on their general principles, which we share, than on the concreteness of their expectations.

Our proposal is accordingly specific. The EDPB should require documented datasheets from those who supply datasets and due diligence from those who re-use them; treat the disregard of machine-readable reservations as a bar to legitimate interest; demand filters that recognise the identifiers and languages actually present in each corpus; weigh the irreversibility of training in favour of stronger ex ante safeguards; confine the analogy with GC and Others to cases where those safeguards are effective; and connect its criteria with the documentation duties of the AI Act. Expectations of this kind would protect data subjects on both sides of the Atlantic and would offer the authorities and legislators of Latin America, currently redesigning their own institutions, a reference they can adopt with confidence. We remain available to the EDPB for any further exchange.

References

- EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted 7 July 2026.
- EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted 17 December 2024.
- EDPB, Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted 8 October 2024.
- EDPB, Report of the work undertaken by the ChatGPT Taskforce, 23 May 2024.
- Autoriteit Persoonsgegevens (Netherlands), decision imposing a fine of €30.5 million on Clearview AI, September 2024.
- Garante per la protezione dei dati personali (Italy), decision imposing a fine of €15 million on OpenAI concerning ChatGPT, December 2024; and Web scraping ed intelligenza artificiale generativa: nota informativa e possibili azioni di contrasto, decision No. 329 of 20 May 2024.
- Joint statement on data scraping and the protection of privacy (24 August 2023) and Concluding joint statement on data scraping and the protection of privacy (28 October 2024), issued by members of the Global Privacy Assembly's International Enforcement Cooperation Working Group.
- CNIL, Recommendations on the development of AI systems: legitimate interest and web scraping ('moissonnage'), 19 June 2025.
- CJEU, C-136/17, GC and Others, judgment of 24 September 2019; C-252/21, Meta Platforms and Others, judgment of 4 July 2023.
- Ley Federal de Protección de Datos Personales en Posesión de los Particulares (Mexico), Diario Oficial de la Federación, 20 March 2025.

- Regulation (EU) 2016/679 (GDPR); Regulation (EU) 2024/1689 (AI Act); Directive (EU) 2019/790 (DSM Directive).