ICT4Law&Forensics
Legal Informatics & Digital Forensics Laboratory
DIEE – Department of Electrical and Electronic Engineering
UNIVERSITY OF CAGLIARI

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

Contribution to the public consultation on Guidelines 03/2026

ICT4Law&Forensics – University of Cagliari Academic contribution | Updated on 30 October 2026

TABLE OF CONTENTS / SOMMARIO

• **Purpose and approach** *[Introductory assessment]*
1. Comparability of processing activities and reliance on GC and Others (C-136/17) *[Paras 67–77]*
2. Foreseeability and intentionality in processing special categories of personal data *[Paras 68–75]*
3. Reasonable expectations and platform choices *[Paras 62–64, Ex. 4–5]*
4. The scope of the balancing assessment in Example 7 *[Para 66, Ex. 7]*
5. Sufficiently specified purposes and evidence of necessity *[Paras 23, 36–38, 48–52]*
6. Coordination between Articles 11 and 12 and operationalizing the right to object *[Paras 29, 33, 60, 66(f)]*
7. Dataset erasure, model interventions, and output suppression *[Paras 60, 74(ii)–(iv)]*
8. Transparency and proportionate assessment of disproportionate effort *[Paras 24–34, Ex. 2]*
9. Legal function of mandatory obligations versus risk-mitigating safeguards *[Paras 34–38, 65–66]*
10. Dataset provenance and supply chain due diligence *[Paras 3–4, 16–20, 32]*
11. Technical signals and legally relevant expressions of choice *[Paras 37, 45, 64, 66]*
12. Processing of criminal offence data, accuracy, and retention *[Paras 39–42, 66(d), 70–72]*
• **Requests for the final version** *[Policy & normative requests]*
• **Key references** *[Statutory & case-law authorities]*

Purpose and approach

This contribution concerns Guidelines 03/2026 on web scraping in the context of generative AI, adopted on 7 July 2026. It supports the development of common guidance enabling compliance with Regulation (EU) 2016/679 to be assessed and demonstrated. It proposes amendments to passages in which technical difficulties associated with processing could be interpreted as relaxing the legal conditions for lawful processing.

The principal request is for a clearer distinction between the lawfulness of processing, the measures necessary to meet the applicable obligations, and the evidence through which controllers must demonstrate compliance. Accountability requires a contextual assessment that reflects the state of the art, but does not allow controllers to replace the conditions laid down by the GDPR with their own technical assessment. The following proposals seek to make data subjects' rights effective while providing developers and dataset suppliers with predictable compliance requirements.

References to paragraphs and examples relate to the consultation draft. The proposed amendments are requests for interpretative clarification or additions to the text; they do not describe further obligations already introduced by the Guidelines.

1. Comparability of processing activities and reliance on GC and Others (C-136/17)

Paragraphs 67–77, particularly 70–74

One of the most sensitive parts of the draft concerns its reliance on the judgment in GC and Others, C-136/17, to address the incidental collection of special categories of personal data through web scraping. While paragraph 72 states that this does not create a general exemption from Articles 9 and 10, the wording of paragraph 74 leaves uncertainty regarding how the "relevant similarities" referred to in paragraph 73(a) are to be established in practice.

In GC and Others (paragraphs 45–48), the Court examined search engine activity, calibrating obligations specifically to the referencing of web pages and the display of links in response to name-based queries. The functional rationale justifying that calibration cannot be automatically transposed to generative AI. To avoid an overly broad or ungrounded analogy, the Guidelines should explicitly differentiate between distinct processing operations, including: (i) collection for web indexing and source referencing; (ii) collection for pretraining or fine-tuning model parameters; and (iii) the dynamic retrieval or querying of external sources during real-time system operation. A single generative AI service may encompass several of these operations, each requiring separate legal assessment.

Furthermore, while EDPB–EDPS Joint Opinion 2/2026 (paragraphs 46–52) notes the potential utility of a tailored statutory regime for incidental processing, the Council's negotiating mandate on the Digital Omnibus (ST 14756/26, 22 October 2026, Article 9(2)(k)) constitutes a proposed legislative amendment that is not yet in force. The ongoing legislative process cannot serve as an interpretative surrogate to broaden the existing derogations under Article 9(2) GDPR in advance.

Proposed amendment. Supplement paragraph 73(a) with verifiable functional criteria to evaluate the comparability of specific processing operations—expressly distinguishing indexing from training and real-time retrieval—and clarify in paragraph 74 that adopting technical safeguards does not create an autonomous derogation under Article 9(2).

2. Foreseeability and intentionality in processing special categories of personal data

Paragraphs 68–75 and footnote 44

The classification of processing as incidental and residual cannot depend solely on the absence of intent on the controller's part, as Article 9(1) imposes an objective prohibition. Equally, however, the mere foreseeability of encountering sensitive data on the open web does not automatically render their processing intentional or non-residual. As footnote 44 recognises, a degree of foreseeable risk is inherent to large-scale collection and may be compatible with residual processing provided robust safeguards operate effectively.

The assessment might therefore focus on the relationship between the concrete purpose, source selection, collection selectivity, and the preventative measures deployed. A clear distinction is needed between: (a) intentionally selecting sources specifically because they contain special category data (e.g., scraping dedicated medical patient forums or political discussion groups); and (b) the limited, non-targeted ingestion of sensitive data from heterogeneous sources despite effective pre-filtering and prompt scrubbing. Foreseeability should govern the standard of diligence and accountability required, rather than operating as an automatic disqualifier of residual processing. Finally, paragraph 74(i) should confirm that while children's vulnerability warrants heightened protection, it does not assimilate children's data to Article 9 special categories per se.

Proposed amendment. Define residual processing through a contextual assessment taking into account source selectivity, foreseeability, and verifiable safeguards. Clarify that foreseeable exposure demands enhanced diligence rather than establishing intent, and provide clear protocols for promptly isolating and purging residual sensitive data.

3. Reasonable expectations and platform choices

Paragraphs 62–64 and Examples 4 and 5

Paragraph 63 correctly states that fulfilling transparency duties does not automatically establish that processing is reasonably expected by data subjects. However, Example 4 concludes too categorically that users can reasonably expect scraping for AI development where a platform is freely accessible, does not prohibit collection, and informs users of scraping in its general notice.

Mere technical awareness that scraping occurs on the web cannot be equated with a reasonable expectation of subsequent processing for generative AI training. Otherwise, widespread data extraction would become self-justifying through sheer prevalence. Nor can platforms unilaterally determine the balancing test for third-party scrapers solely by amending their terms of service or privacy notices. The context of original publication remains relevant: publishing a CV on an employment platform or posting in a public debate does not, in itself, establish a reasonable expectation of subsequent processing for model training. The position of individuals who have no realistic platform alternatives, or third parties mentioned by users, must be evaluated within their specific context.

Proposed amendment. Revise Example 4 into a multi-factor contextual assessment. Clarify that accessibility, platform notices, and the lack of technical blocks are contextual elements that do not, on their own, establish a reasonable expectation of AI training under Article 6(1)(f).

4. The scope of the balancing assessment in Example 7

Paragraph 66 and Example 7

The assertion that the balancing test may generally be considered satisfied where the conditions in Example 7 are met risks establishing an unjustified presumption of a favourable balancing outcome. The test under Article 6(1)(f) requires a case-specific assessment that cannot be settled categorically in advance.

Example 7 also conflates public availability, copyright exceptions, and data protection rules. The absence of a reservation of rights under Article 4(3) of Directive (EU) 2019/790 may allow the text and data mining (TDM) exception to apply only where the other conditions laid down in the Directive, including lawful access, are met; it neither places works in the public domain nor establishes the lawfulness of personal data processing under the GDPR. Works in the public domain must be distinguished from protected works that may be used under a copyright exception.

Regarding Article 9(2)(e), the judgment in Schrems (C-446/21, paragraphs 79–83) establishes that a public statement about one's sexual orientation does not, by itself, authorise a social network operator to aggregate and analyse other data concerning that person's sexual orientation for personalised advertising. As a distinct and autonomous consideration, the manifest publication of content by a data subject cannot be equated with the manifest publication by third parties whose personal data may be contained in that same content.

Proposed amendment. Clarify that the conditions in Example 7 do not create a presumption of a favourable balancing test outcome. Expressly differentiate copyright TDM conditions from lawful processing under the GDPR, and align the application of Article 9(2)(e) with the precise ruling in C-446/21 while treating third-party data as a distinct legal issue.

5. Sufficiently specified purposes and evidence of necessity

Paragraphs 23, 36–38 and 48–52

For general-purpose AI models, defining the purpose as "developing AI" or "maximising model performance" risks making virtually any accessible data appear necessary. The Guidelines should require controllers to articulate reasonably foreseeable classes of use, intended core capabilities, and data categories without demanding foresight of every future application.

The necessity of processing under Article 6(1)(f) must be substantiated. Depending on the circumstances, representative sampling, source comparisons or performance ablation studies may provide evidence for assessing whether less intrusive alternatives could achieve the defined objective as effectively. These are illustrative methods, not a mandatory or exhaustive set of technical tests. Marginal performance gains do not equate to legal necessity, and practical difficulties in obtaining consent cannot broaden the scope of legitimate interest.

Proposed amendment. Require controllers to substantiate necessity through verifiable evidence—such as comparative evaluations with less intrusive alternatives—distinguishing marginal utility from legal necessity, and clarify that consent hurdles do not expand legitimate interest grounds.

6. Coordination between Articles 11 and 12 and operationalizing the right to object

Paragraphs 29, 33, 60 and 66(f)

The draft requires careful alignment regarding the interplay between Articles 11 and 12 GDPR. While Article 11(2) specifically lists Articles 15 to 20, Article 12(2) applies to requests under Articles 15 to 22. In situations governed by Article 11(2), Article 12(2) allows controllers to refuse to act on grounds of lack of identification where they demonstrate that they are not in a position to identify the data subject, unless the data subject provides additional information enabling identification. This clarification concerns identification and is without prejudice to other grounds for refusing a request under the GDPR.

The Guidelines should clearly distinguish between verifying the identity or entitlement of the requester and locating the relevant data records. Controllers may be able to locate data using URLs, platform usernames or text excerpts without knowing the individual's legal name. While these functional references may enable data to be located, they do not necessarily prove that the requester is the person concerned or authorised to act. Procedures should maintain this distinction. Any additional identity checks must remain necessary and proportionate, in accordance with Article 12(6). Furthermore, paragraph 66(f) must not dilute the statutory right to object under Article 21: under Article 21(1), the controller must cease processing unless it demonstrates compelling legitimate grounds overriding the data subject's interests, rights and freedoms, or processing is necessary for the establishment, exercise or defence of legal claims. This statutory mechanism must not be conflated with voluntary pre-collection opt-outs.

Proposed amendment. Coordinate the guidance under Articles 11(2), 12(2) and 12(6) regarding refusals based on identification, distinguishing record location via functional references from necessary and proportionate identity verification. Reaffirm the full statutory scope of Article 21(1)—including the exception for legal claims—distinguishing it from discretionary pre-scraping signals.

7. Dataset erasure, model interventions, and output suppression

Paragraphs 60 and 74(ii)–(iv)

Effective data subject rights require a clear distinction between: (a) erasing personal data from training datasets; (b) model-level interventions (e.g., machine unlearning, weight updates); and (c) output suppression through filters or other safeguards. Output filtering alone does not demonstrate erasure under Article 17: personal data may remain retrievable from the model despite restrictions on particular outputs. The assessment should establish whether personal data remain in the model and whether the measure achieves the outcome required by the applicable right. Model-level interventions must also be assessed for their actual effectiveness.

Proposed amendment. Differentiate the technical and legal outcomes of dataset scrubbing, model retraining/unlearning, and output filtering, defining clear retention and evidence standards for each layer.

8. Transparency and proportionate assessment of disproportionate effort

Paragraphs 24–34 and Example 2

Relying on Article 14(5)(b) should not follow automatically from the scale of collection. Scale alone does not establish disproportionate effort. Datasets should be evaluated by homogeneous source categories: recent sources with available contact information should not be exempted merely because they are aggregated with historical data lacking contact details. Public source repositories must also be minimised to avoid publishing lists that inadvertently expose sensitive personal URLs.

Proposed amendment. Clarify in Example 2 that collection scale alone does not establish disproportionate effort, requiring granular assessments across homogeneous dataset subsets.

9. Legal function of mandatory obligations versus risk-mitigating safeguards

Paragraphs 34–38 and 65–66

Paragraph 65 correctly distinguishes mandatory GDPR obligations from additional safeguards in the legitimate interest balancing test. However, the subsequent list risks double-counting compliance baseline measures (e.g., baseline storage limitation or carrying out a mandatory DPIA under Article 35) as positive compensatory safeguards. The Guidelines should distinguish the legal function of a measure (as an inherent compliance duty) from its practical efficacy in mitigating risk. Compliance with baseline obligations cannot be credited as an additional voluntary safeguard compensating for weak legal grounds.

Proposed amendment. Restructure the balancing framework to clarify that adhering to baseline legal duties (including mandatory DPIAs) does not constitute an additional voluntary safeguard, while evaluating the practical efficacy of risk-mitigating measures contextually.

10. Dataset provenance and supply chain due diligence

Paragraphs 3–4, 16–20 and 32

Where developers procure pre-scraped third-party datasets, the supplier's independent controller status does not discharge the buyer's duty to ensure the lawfulness of its own processing. Purchasers must conduct proportionate due diligence covering collection provenance, licensing scope, filtering protocols, and mechanisms for executing deletion requests, without imposing unbounded strict liability across upstream actors.

Proposed amendment. Provide a proportionate checklist of provenance checks and verifiable documentation expected of organisations acquiring external datasets for AI development.

11. Technical signals and legally relevant expressions of choice

Paragraphs 37, 45, 64 and 66

The Guidelines should evaluate technical signals according to their content, origin, and legal effects, avoiding a rigid correspondence between specific formats and legal regimes. Robots.txt communicates crawling preferences; it is not an access-authorisation mechanism (RFC 9309, Section 1). Similarly, technical signals may express server-level management preferences, copyright reservations under Article 4(3) of Directive (EU) 2019/790, or data protection objections under Article 21 GDPR. The same technical format may serve more than one function. Whether a signal constitutes a reservation under Article 4(3) or an objection under Article 21 depends on the applicable legal conditions and the authority of the person expressing it.

The lack of a technical signal does not imply GDPR consent, nor can an operator's signal determine the data protection rights of third parties mentioned in hosted content. Guidance is needed to classify how different signals operate in practice and how their legal significance should be evaluated across access, copyright, and data protection rules.

Proposed amendment. Classify technical signals according to their origin, content, and legal effects across access, copyright, and privacy, clarifying that technical formats do not, by themselves, establish consent or authorise a website operator to waive third parties' data protection rights.

12. Processing of criminal offence data, accuracy, and retention

Paragraphs 39–42, 66(d), 70–72

Where scraped reporting contains personal data falling within Article 10 GDPR, the conditions of that provision must be assessed separately from the derogations under Article 9. Although neither Latvijas Republikas Saeima (C-439/19, paragraphs 120–129) nor Endemol Shine Finland (C-740/22, paragraphs 48–58) directly addressed AI model training—the former concerning public access to penalty-point data and its commercial re-use, and the latter concerning oral disclosure of conviction records from court registries—these judgments support the need to assess the lawfulness of disclosure and subsequent processing separately, rather than treating public accessibility as sufficient.

Furthermore, generative outputs can produce contextual inaccuracies by misrepresenting historic allegations as current convictions or generating confabulations affecting namesakes. Verification mechanisms and justified retention criteria are essential: retention criteria should be defined and justified for raw scrapes, curated datasets, and system logs, and controllers must evaluate anonymity claims strictly in line with EDPB Opinion 28/2024, recognising that pseudonymous and synthetic data are not inherently anonymous.

Proposed amendment. Incorporate a dedicated section on Article 10 GDPR integrating the principles of C-439/19 and C-740/22, providing concrete safeguards against contextual inaccuracy and setting distinct retention standards across dataset lifecycle phases.

Requests for the final version

We propose that the final version of the Guidelines provide an explicit functional justification for any reliance on C-136/17, clarify the relationship between foreseeability, intent, and safeguards under Article 9, refine the balancing test in Examples 4 and 7, and establish practical standards for exercising rights under Articles 11, 12, and 21 using functional identifiers.

Legal certainty requires distinguishing statutory obligations from non-binding best practices. Evolving technical state-of-the-art must guide the implementation of safeguards without eroding the protective architecture of Regulation (EU) 2016/679.

Key references

• EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted on 7 July 2026.
• Regulation (EU) 2016/679 (GDPR), particularly Articles 5, 6, 9, 10, 11, 12, 14, 17, 21, 24, 25, 35.
• Court of Justice, Grand Chamber, 24 September 2019, GC and Others, C-136/17, ECLI:EU:C:2019:773.
• Court of Justice, 4 October 2024, Schrems, C-446/21, ECLI:EU:C:2024:834.
• Court of Justice, Grand Chamber, 22 June 2021, Latvijas Republikas Saeima, C-439/19, ECLI:EU:C:2021:504.
• Court of Justice, 7 March 2024, Endemol Shine Finland, C-740/22, ECLI:EU:C:2024:216.
• EDPB–EDPS, Joint Opinion 2/2026 on the Digital Omnibus, adopted on 10 February 2026.
• Council of the European Union, ST 14756/26, 22 October 2026, negotiating mandate on the Digital Omnibus.
• Directive (EU) 2019/790 (DSM Directive), particularly Articles 3 and 4 on text and data mining.
• IETF, RFC 9309, Robots Exclusion Protocol, September 2022.
• EDPB, Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024.