Contribution to the EDPB Public Consultation

on the Guidelines 03/2026 on Web Scraping in the Context of Generative AI (v1.0, 07 July 2026)

Observations from the perspective of a Lawyer, Data Protection Officer and Registered Journalist:
Submitted by: Avv. Francesco Luongo — Cassazionista, Lawyer, Data Protection Officer, Registered Journalist (ODG Campania)
Contribution category: DPO / Individual
Country: Italy — www.linkedin.com/in/francesco-luongo
Sector: Professional and technical activities; media; consumer protection

1. Premise and Position of the Contributor

This contribution is submitted in a personal and professional capacity, at the intersection of three roles: appointed Data Protection Officer of a professional body governed by public law, cassation-level lawyer active in consumer protection litigation, and registered journalist. None of these roles is represented in an official capacity by this submission, which reflects the contributor's own analysis.

The observations below draw on operational experience with three recurring situations that the draft Guidelines, being focused on private-sector controllers scraping open web content for AI training, do not directly address: the scraping of journalistic and editorial content authored under a professional publishing framework; the scraping of statutorily-mandated public professional registers; and the practical difficulty faced by individual data subjects — and the utility of consumer associations — in exercising rights against large-scale, indiscriminate scraping.

This standpoint is relevant to the consultation because the reasonable-expectations and balancing-test analysis in the Guidelines is built primarily around ordinary social-media or forum publication, and does not yet fully account for content published under an institutional or statutory framework that pursues purposes other than open, unrestricted reuse.

2. General Assessment

The Board's decision to address, for the first time in a dedicated document, the specific compliance challenges raised by web scraping for generative AI training is welcomed. The structured treatment of the controller/processor qualification, the purpose-limitation and transparency principles, the data-minimisation measures, and in particular the three-condition legitimate-interest test with its detailed balancing-test guidance, provides a solid and practical basis for controllers to organise their compliance.

The observations that follow are offered to refine, rather than to contest, that framework, with particular attention to categories of scraped content — journalistic and editorial material, and statutorily public professional registers — whose original publication purpose differs materially from voluntary, unrestricted self-publication, and to the practical means by which data subjects can exercise their rights collectively rather than individually.

3. Journalistic and Editorial Content and the "Reasonable Expectations" Test

Paragraphs 62–64 correctly identify the data subjects' reasonable expectations as a central element of the balancing test, and list several relevant factors: the nature of the source website, the type of publication and its degree of public accessibility, technical restrictions imposed by the source, the relationship between

the data subject and the controller, and the characteristics of the data subject. Examples 4 and 5 illustrate the point by reference to a content-sharing platform that does, or does not, technically restrict scraping.

This analysis, calibrated on ordinary user-generated content, does not fully capture the position of journalistic and editorial content published under a professional framework. A journalist's byline, appearing on an investigative article under the statutory and professional accountability framework applicable to journalistic activity — including Article 2 of Law No. 69/1963 on the rights and duties of journalists and the press-publication framework under Law No. 47/1948 — is personal data published for a transparency and accountability purpose towards readers — so that responsibility for the piece can be traced — and not for the purpose of enabling that content, or the journalist's identity, to be repurposed as training material for a competing or derivative generative output. The absence of a robots.txt restriction on a newspaper's website, which may itself reflect resource constraints rather than a deliberate policy choice, should not automatically be read as an indication that content published under this professional framework carries the same low expectation of reuse as ordinary self-published social-media content.

Recommendation: the Board should consider adding, to the non-exhaustive list of factors at paragraph 64, the professional or institutional purpose for which the content was originally published, distinguishing content published to fulfil a transparency, accountability or statutory-disclosure function from content voluntarily and unrestrictedly published by the data subject for general circulation.

4. Facilitating the Collective Exercise of Data Subjects' Rights

Paragraph 60 candidly acknowledges that, in practice, it is difficult for individual data subjects to exercise their rights and to oppose scraping, particularly where data are scraped by third parties with whom they have no direct relationship. Among the mitigating measures listed at paragraph 66, point (f) proposes that controllers facilitate the exercise of individuals' rights, including through an opt-out list allowing data subjects to object to collection of their data.

This measure, while valuable, still places the initiative and the transactional cost on each individual data subject. Under Italian law, consumer associations meeting the representativeness requirements of Article 137 of the Consumer Code, and empowered to bring collective redress actions under Articles 840-bis et seq. of the Code of Civil Procedure, already perform this aggregating function in other domains of consumer protection, including data-related grievances. Recognising an equivalent role in the specific context of web scraping for generative AI — for instance through collective opt-out registries maintained or endorsed by recognised consumer associations, to which individual data subjects could adhere once rather than object to each individual scraping controller — would reduce the very effort asymmetry the Guidelines identify as a practical obstacle and would strengthen the balancing-test analysis by offering a practical mitigating safeguard. Where such mechanisms are designed as ex ante exclusion lists or source-filtering criteria, they may also support data minimisation and the necessity assessment by enabling the controller to achieve the same development purpose through less individually burdensome means.

Recommendation: the Board should consider explicitly recognising collective or associative facilitation of opt-out and objection mechanisms — consistent with existing consumer-protection collective redress frameworks — as a relevant, and in some respects preferable, form of the mitigating measure already described at paragraph 66(f).

5. Statutorily Public Professional Registers and the Limits of "Publicly Accessible" as a Marker of Reduced Expectation

Drawing on direct experience as Data Protection Officer of a professional body governed by public law, whose online register of members (Albo) is, by design, publicly searchable, the contributor notes a further nuance relevant to the reasonable-expectations analysis. Professional registers of lawyers, and comparable registers for other regulated professions, are made publicly accessible pursuant to a specific statutory transparency purpose: enabling members of the public to verify a professional's registration status before engaging their services, a purpose closely tied to consumer protection rather than to any implied consent to secondary reuse.

Such bodies, moreover, are frequently non-economic public entities with limited technical capacity to implement restrictions such as robots.txt directives or CAPTCHA challenges on their registers, a constraint that should not, by omission, be read as a signal of indifference to bulk collection within the meaning of paragraph 64's discussion of restrictions imposed by the scraped website.

Recommendation: alongside Examples 4 and 5, the Guidelines should clarify that data made public pursuant to a specific statutory or regulatory transparency obligation — as opposed to voluntary, unrestricted publication by the data subject — should generally weigh towards a finding that data subjects cannot reasonably expect that data to be scraped for generative AI training purposes, and that the absence of technical anti-scraping measures on such registers, standing alone, should not be treated as displacing that conclusion where the controller is a small public-law body with limited technical means.

6. Incidental Special-Category Data in Professional and Legal-Commentary Content

Legal and journalistic commentary on court proceedings, professional case studies and similar editorial content may incidentally reveal special-category data concerning third parties — health conditions, addiction, criminal proceedings — even where the primary subject matter is a matter of legitimate public interest. This is structurally analogous to the search-engine scenario examined by the Court of Justice in GC & Others (C-136/17), which the Guidelines already discuss at paragraphs 70–77 in relation to the incidental and residual collection of special categories of personal data "within the framework of [the controller's] responsibilities, powers and capabilities."

Recommendation: the Board should confirm expressly that the responsibilities, powers and capabilities test set out at paragraphs 72–74 can also be relevant to the large-scale, automated scraping of professional, legal-commentary and journalistic content, subject to a case-by-case assessment, and that the output-filtering measures described at points (iii)–(iv) of paragraph 74 should be calibrated specifically to prevent AI systems trained on such content from re-surfacing incidentally-collected special-category data concerning third parties who are not the subject of the original public-interest reporting.

7. Aligning Documentation with Existing Organisational Compliance Models

The mitigating measures and documentation practices described throughout Section 3.2.3, and in particular the transparency safeguards at paragraph 66(e), will in practice often be produced and maintained by the same compliance functions responsible for an organisation's broader governance model — including, where applicable, wider organisational accountability and risk-management records such as, for Italian entities, the organisational and management models adopted under Legislative Decree 231/2001, as well as AI Act documentation where the controller is also subject to AI-specific obligations.

Recommendation: the Board should consider clarifying that documentation produced to demonstrate compliance with the data-minimisation and legitimate-interest safeguards described in these Guidelines can, and should, be integrated into a single, coherent accountability and governance record, where compatible with each framework's specific requirements, rather than maintained as a parallel and

potentially inconsistent documentation set, particularly for smaller controllers subject to organisational-liability, data-protection and AI-specific compliance obligations.

8. Conclusions

The contributor supports the Board's initiative to provide dedicated guidance on web scraping for generative AI training and considers the proposed framework, in particular the detailed balancing-test methodology, a valuable and largely workable tool. The observations above are intended to ensure that the final Guidelines properly account for content published under a professional, editorial or statutory framework, and to strengthen the practical means by which data subjects — individually or collectively — can exercise their rights.

In summary, it is respectfully suggested that the Board consider: (i) adding the original professional or institutional publication purpose as a factor in the reasonable-expectations analysis; (ii) recognising collective, associative facilitation of opt-out and objection mechanisms as a valid and often preferable mitigating measure; (iii) clarifying that data made public pursuant to a statutory transparency obligation should generally weigh against a finding of reduced reasonable expectation, irrespective of the technical anti-scraping capacity of the hosting body; (iv) confirming that the responsibilities, powers and capabilities test from GC & Others can also be relevant to professional and journalistic content, subject to case-by-case assessment; and (v) encouraging the integration of scraping-related documentation into existing organisational compliance models, where compatible with the specific requirements of each framework.

The contributor remains available to provide further detail or practical examples on any of the points raised, and thanks the Board for the opportunity to contribute to this consultation.

Yours faithfully,
Avv. Francesco Luongo