Position Paper

EDPB Guidelines 03/2026 on Web Scraping in the Context of Generative AI: Numeum calls for a workable and risk-based framework

Numeum response to the public consultation, September 2026

Executive summary

Numeum, the French trade association for the digital sector, represents 2,500 technology companies, from innovative SMEs to large international players, covering the full digital value chain, including AI, cloud services, data management, cybersecurity, and digital infrastructures. Our members are directly affected by the conditions under which data may lawfully be collected online to train and fine-tune generative AI models.

Numeum welcomes several clarifications in the draft, which provide useful foundational certainty for the sector: the explicit recognition of legitimate interest under Article 6(1)(f) as a valid basis for scraping to train AI models, the confirmation in paragraph 36 that data minimisation does not prohibit training on large volumes of data, the acknowledgement in paragraphs 25 to 28 that the disproportionate effort exception under Article 14(5)(b) GDPR applies to large scale web scraping, the confirmation in paragraph 29 that Article 11 GDPR protects controllers from acquiring additional identifying information solely to respond to data subject requests.

Several expectations set out in the draft also cannot be met with current state of the art technologies, and others transpose assumptions drawn from structured database processing onto AI training. Numeum calls on the EDPB to calibrate these requirements to the technical reality of AI development, and to ensure that the Guidelines remain compatible with the outcome of the Digital Omnibus negotiations.

1) Ground the Guidelines in how AI models are built

➢ Describe the AI development lifecycle

Several of the expectations set out in the draft appear to transpose assumptions drawn from structured database processing onto AI training. A short preliminary section describing how generative AI models are developed, from broad ingestion through to deployment, would help anchor the subsequent analysis in technical reality and avoid requirements that cannot be mapped onto the development pipeline.

Recommendation: The Guidelines should include an introductory description of the AI development lifecycle and identify, for each stage, which obligations attach and which do not.

➢ Acknowledge untargeted crawling as a technical reality

The Guidelines implicitly favour targeted scraping with restrictive collection criteria over untargeted crawling, and paragraph 52 suggests that narrowing collection criteria may be key in establishing that the necessary conditions under article 6(1)(f) are met. This does not reflect the architecture of general-purpose and foundation models, which depend on broad and diverse data ingestion to acquire linguistic and contextual capability.

Recommendation: The Guidelines should explicitly acknowledge that broad, untargeted crawling is a standard technical reality in the development of general-purpose AI models, without treating the restrictiveness of the collection perimeter as determinative of necessity. This recognition should not, however, amount to a general presumption. Untargeted crawling may be relevant to the development of certain general-purpose models without being necessary or proportionate in every case.

Necessity and proportionality should continue to be assessed by reference to the purpose pursued, the scale of the collection, the sources relied on and the intended use of the data, since these vary considerably from one development scenario to another. The Guidelines should also recognise targeted source selection, documented provenance, regular updating and filtering as safeguards capable of supporting that assessment.

2) Make the transparency framework operable at scale

➢ Apply the disproportionate effort exception on the basis of scale

Three elements of the transparency section are not workable as drafted.

First, Article 14(5)(b) GDPR relieves controllers of the obligation to inform individuals whose data was not obtained from them where doing so proves impossible or would involve a disproportionate effort. This is the situation web scraping typically creates. The statement in paragraph 27 that this exception should not be relied upon routinely creates ambiguity. Whether the exception applies should be assessed on the basis of the specific circumstances of the processing.

Paragraph 28 of the Guidelines identifies relevant factors for this assessment, including the number of data subjects concerned, the age of the data and the safeguards adopted by the controller. These should determine whether providing the information would be impossible or would involve a disproportionate effort in the circumstances of the case. The frequency with which a controller relies on the exception should not, in itself, determine whether Article 14(5)(b) applies. The reference in paragraph 27 to the exception not being "relied upon routinely" should therefore be removed or clarified. In line with paragraph 29 and Article 11 GDPR, paragraph 28 should also make clear that relying on Article 14(5)(b) does not require controllers to quantify or identify the number of identifiable individuals present across unstructured web corpora.

Second, paragraph 30 suggests that privacy notices should indicate the types of sources excluded from the collection and the reasons for their exclusion, while paragraph 31 expects controllers to provide domain names, searchable URLs and precise collection dates. Article 14(2)(f) requires the source from which the data originate to be disclosed. It does not require controllers to account for sources they have chosen not to collect, and such negative disclosures would not assist data subjects in understanding the processing or exercising their rights.

Third, the Guidelines should also take into account the transparency framework already established by the EU legislator for GPAI models under Article 53(1)(d) of the AI Act. The Commission's template adopts a scalable approach to source transparency, requiring information on the crawlers used, the collection period, the type of content collected and the main domains scraped, rather than requiring an exhaustive URL-by-URL disclosure. The framework also provides for reduced requirements for SMEs, reflecting the importance of proportionality where training data are collected at very large scale. Although the two instruments pursue different objectives, this framework provides a relevant benchmark for proportionate source transparency in the context of AI training.

Recommendation: The EDPB should confirm that the Article 14(5)(b) exception applies whenever the factual conditions are met, and remove the routinely restriction. The expectation in paragraph 30 to list excluded sources should be removed, and paragraph

31 should be revised to provide that documenting sources at the category or domain type level is sufficient.

The EDPB should clarify that Article 14 does not systematically require controllers to identify individual URLs or other granular source-level information where doing so would be technically impracticable or disproportionate. Aggregated information at domain or source category level should be considered sufficient where it enables data subjects to understand the sources and nature of the processing and to exercise their rights effectively. The Guidelines should also recognise that source transparency may be limited where necessary to protect business secrecy, proprietary methodologies, security arrangements (including protection against targeted data poisoning and extraction attacks) and competitively sensitive information.

➢ Preserve the confidentiality of data protection impact assessments

Paragraph 34 suggests making data protection impact assessments publicly available. This goes beyond Article 35 GDPR, which imposes no obligation of publication, and would expose commercial confidentiality and system security information, while undermining the confidentiality of the exchanges between controllers and their lead supervisory authority under Articles 35 and 36 GDPR.

Recommendation: The suggestion to publish DPIAs should be deleted. DPIAs should remain confidential and available to supervisory authorities upon request, transparency towards data subjects being delivered through privacy policies and dedicated AI notices.

3) Clarify opt-out mechanisms and the boundary with copyright law

➢ Clarify the status of technical opt-out signals

Paragraph 37 would, for the first time, treat the failure to observe technical opt-out tools such as robots.txt as a compliance failure under the GDPR. These tools, however, have not been designed for the purpose of personal data protection and the Guidelines do not acknowledge their limitations. It also raises a question of articulation with copyright law, where Article 4 of Directive 2019/790 (CDSM directive) already governs the reservation of rights for text and data mining, including for commercial purposes.

Paragraph 66(f) adds to this practical difficulty by suggesting the creation of opt-out lists allowing individuals to object before collection, which would require controllers to identify individuals across sites in advance, that is, the very outcome the Guidelines seek to avoid elsewhere under Article 11 GDPR. Paragraph 66(c) raises a comparable difficulty by suggesting the exclusion of data from freely accessible websites where data subjects

cannot be considered reasonably aware of that accessibility. A third party crawler has no means of assessing what individuals believe about the visibility of their own content. Accessibility should be assessed against objective and technically verifiable criteria, namely whether content is publicly reachable without authentication or technical access controls.

Recommendation: The expectation of centralised opt-out lists in paragraph 66(f) should be removed in favour of existing standard, proven and decentralised machine-readable technical signals and paragraph 66(c) should rely on objective technical accessibility rather than on the awareness of data subjects. In parallel, the EDPB should acknowledge the current limitations of these mechanisms, encourage their standardisation, and clarify the articulation between the data protection assessment and the reservation of rights regime under copyright law. The Guidelines should further clarify that technical signals such as robots.txt must be correctly implemented, standardised and machine readable before non-compliance with them can give rise to consequences under the GDPR. Where website operators publish contradictory, malformed, inaccessible or out of date opt-out signals, this should be taken into account in assessing whether a crawler has failed to respect them.

- **Separate copyright compliance from the data protection assessment**

Example 7 describes an organisation whose processing would satisfy the balancing test. Among the factors leading to that conclusion, the Guidelines list the exclusion of copyright protected content, the organisation relying only on public domain material or on content for which rightholders have not reserved their rights against text and data mining under Directive (EU) 2019/790.

Copyright compliance thereby becomes a factor in establishing a legal basis under Article 6(1)(f). These are, however, two distinct regimes protecting distinct interests: the GDPR protects the individuals whose data is processed, while copyright protects rightholders in respect of works. The two do not overlap in practice either, since personal data frequently appears in content that is free of copyright, and protected content frequently contains no personal data at all. Making one condition upon the other creates legal uncertainty for controllers without any corresponding benefit for data subjects, whose protection does not depend on the copyright status of the material in which their data appears.

Copyright also falls outside the remit of the EDPB. By embedding it in the balancing test, the Guidelines produce normative effects in a field that is governed by other authorities and other courts.

Recommendation: Example 7 should be amended to strictly separate copyright compliance from the data protection assessment.

4) Introduce a reasonable steps standard for special category data

Numeum welcomes the application, in paragraphs 70 to 74, of the framework established by the Court of Justice in GC and Others (C-136/17), which assesses obligations within the controller's responsibilities, powers and capabilities[1]. Paragraph 74(ii) requires that incidentally collected special category data be deleted from datasets immediately. Pre-scrubbing the open internet to guarantee the absolute absence of such data is not achievable with current state of the art technologies. Unlike structured identifiers such as telephone numbers or email addresses, special category data is inherently contextual, so that real-time filtering across billions of web pages produces high false positive rates, and scraped text is rapidly tokenised and fragmented across computing clusters, which makes immediate isolation and extraction impracticable. The Guidelines simultaneously require state of the art output filters during deployment, imposing a double compliance burden that demands technological perfection at both stages independently.

Recommendation: The EDPB should align paragraph 74 with the GC and Others[2] standard by introducing the concept of reasonable and proportionate steps into paragraph 74, shifting from an absolute deletion guarantee to a workable risk mitigation framework. The Guidelines should recognise that a holistic combination of reasonable ingestion phase safeguards and deployment phase output filtering constitutes sufficient mitigation. Security by design should be the governing principle, supported by documented technical and organisational measures rather than by an unattainable guarantee of complete exclusion.

5) Ensure coherence with the Digital Omnibus

The Guidelines are being finalised in parallel with the negotiations on the Digital Omnibus, which would address the same notions at legislative level, in particular the legal basis for AI development and the treatment of incidentally collected sensitive data.

Recommendation: The EDPB should ensure that Guidelines 03/2026 remain sufficiently flexible, proportionate and risk based so as not to contradict the outcome of the Digital Omnibus negotiations and should commit to reviewing them once the co-legislators reach political agreement, affording stakeholders a formal opportunity to make further submissions once that legislative process has concluded.

[1] (CJUE, grande chambre, 24 septembre 2019, GC e.a. c. CNIL, C-136/17, ECLI:EU:C:2019:773, points 45 à 48)
[2] Ibid

Conclusion

Numeum supports the EDPB objective of clarifying the conditions under which data may be collected online for the development of generative AI. The draft contains clarifications that European companies have been waiting for, and these should be preserved.

To deliver on that objective, the Guidelines must be calibrated to what is technically achievable at the scale at which AI models are built, and must avoid importing requirements designed for structured database processing or belonging to other legal regimes. Numeum stands ready to contribute technical input and concrete examples, and to engage constructively with the EDPB to ensure that these Guidelines support both a high level of protection for individuals and the conditions for responsible innovation in Europe.