Response to the public consultation on Guidelines 03/2026 on web scraping in the context of generative AI

Submitted by: Maja Kurek, AIGP, CIPM, CIPP/E · RuleBridge Advisory
Contact: Zagreb, Croatia · office@rulebridge.eu · rulebridge.eu

Scope and principal recommendation

A collector may register an exclusion that never reaches a developer's training copy; a model user may detect a disclosure without a route to prevent its recurrence. The final Guidelines should make these dependencies explicit: which material and activity a safeguard covers, what supports reliance on it, who can give a finding effect, and which changes require reconsideration.

This response offers five targeted additions across collection, dataset reuse, development and use, in paragraph order. References are to the consultation version adopted on 7 July 2026 [1]. The proposals support existing processes rather than a new mandatory template; they do not create a lawful basis, relax Article 9 or replace the factual allocation of GDPR responsibilities. The draft itself observes, in paragraph 14, that meeting the accountability principle over such datasets is not obvious; these additions are offered as part of the answer to that difficulty.

1. Connect a supplier's safeguard to the material and use it covers

Suggested placement: after paragraph 20; related paragraphs 32 and 66(f).

Paragraphs 16–20 distinguish processing on behalf of another entity, separate controllership and joint controllership. The information needed to rely on a safeguard nevertheless has to cross those boundaries. A collection policy describes an intended practice; it does not establish that the delivered version reflects that policy, or that later exclusions reached copies already supplied. CNIL's existing recommendations cover checks on reused datasets, recipient notifications, downstream rights arrangements and dataset-version records. The addition below connects those existing practices to the material actually processed and to evidence of the action taken; it does not propose an unlimited upstream audit [2].

Proposed wording

> Where a controller relies on a safeguard performed by another entity, it should identify the information needed to establish the safeguard's applicability to its own processing. Relevant information may include the collection period, source information and relevant source conditions at collection, collection and exclusion criteria, the exclusions actually applied, a suitable dataset or batch reference, and material limitations or unresolved findings. The level of detail should be proportionate to the risk and the safeguard relied upon. A general supplier assurance does not establish that the delivered material is covered.
>
> Where an advance opt-out mechanism under paragraph 66(f) is relied upon, its scope should be clear: the entities, collection and subsequent processing it covers, when an exclusion takes effect, and how it reaches the covered datasets. If the protection is intended to operate across entities, the arrangements should provide for the necessary exclusion information to reach each covered operation. They should not imply that registration with one entity automatically binds every independent controller, or require circulation of an entire register where a narrower exclusion instruction is sufficient.

This proposed safeguard arrangement is distinct from the exercise of Article 21 rights and from Article 19 notifications following rectification, erasure or restriction. Article 19 has specified conditions and an impossibility/disproportionate-effort qualification; it is not a general duty to forward every objection to every actor in an AI supply chain. Processor assistance remains governed by Article 28. None of these distinctions permits a controller to disregard a request that concerns its own processing [3].

2. Return material changes to the legitimate-interest assessment

Suggested placement: after paragraph 66 in section 3.2.3; anchored in paragraph 64.

Paragraph 64 recognises that a model's impact may change over time. Paragraph 76 addresses regular verification of measures. These are connected but different questions: a measure may still perform as designed while the processing, affected population or expected impact no longer matches the earlier assessment. This calls for a route from changed facts to reconsideration of the assessment and of what processing may continue.

Proposed wording

> The controller should document the date, processing scope and material factual assumptions of its legitimate-interest assessment. It should identify changes capable of affecting the legitimate interest, necessity or balance assessed, and arrange for relevant information to reach the function responsible for reconsidering the assessment. Examples include material changes in source conditions, data or affected populations, the model's capabilities or use, and the scope or effectiveness of relied-upon safeguards. Objections and findings may also reveal a wider issue; their significance should be assessed rather than made dependent on a numerical threshold.
>
> For each material change, the controller should determine which parts of the earlier reasoning remain applicable, what further assessment is needed, and what collection or reuse may lawfully proceed meanwhile. The record should identify responsibility for obtaining the necessary information, the next action and its timing. A scheduled review date is not permission to disregard an earlier relevant change. Individual rights and applicable restrictions on processing must be addressed without waiting for the wider review.

A new restriction on a source requires careful temporal treatment. Its implications for a future crawl and for continued reuse of previously obtained material should be considered separately. It does not, merely by appearing later, establish that an earlier collection was unlawful. Equally, a decision made before the restriction is not permission for every later operation. A website's technical signal, a rights reservation and a data subject's objection should not be treated as interchangeable legal acts.

Where an Article 21(1) objection is being assessed, the Article 18(1)(d) right to restriction must be respected. A rise in aggregate opt-outs can prompt a systemic review; it must never become a threshold below which individual rights receive no response.

Related field evidence illustrates why this route needs an owner: in the author's study of a post-deployment AI arrangement, technical capability and receipt of change information were not equivalent to assigned action or an organisational response [4].

3. Make removal and output-monitoring duties executable

Suggested placement: additions to paragraphs 74(ii) and 74(iv); cross-reference from paragraph 33.

Removal: follow the affected material, not only the request. A finding can reach the collector after delivery but before a transformed copy enters training. Closing a ticket at the collector does not establish deletion at the developer. The clarification should preserve the draft's prompt-deletion requirement, not create a supplier-response waiting period.

Proposed addition to paragraph 74(ii)

> Deletion arrangements should address the affected copies and versions within the controller's processing, including cleaned, transformed and staged material. Where another function or entity must act, the controller should identify the action required, provide a timely route to the party able to perform it, and obtain evidence of completion appropriate to its responsibilities. Relevant obligations to notify recipients and obtain processor assistance remain applicable. Internal reassignment or a supplier handoff does not restart a statutory response period.

Where deletion is required under paragraph 74(ii), the identified material should not enter further training while removal is being coordinated. Preventing further use does not itself provide a legal basis for retention. Records should distinguish exclusion from a future batch, deletion from a stored dataset, action concerning

an already trained model, and the response to the data subject. Evidence of one outcome should not be presented as evidence of another.

In particular, corpus-level deletion or suppression of an output should not be described as removal from model parameters without evidence supporting that claim. Article 9 conditions, the draft's narrowly framed residual-data discussion and any required action concerning a trained model must be assessed in their respective scope, not resolved by relabelling a dataset operation.

Monitoring: identify the information and response the model user needs. Paragraph 74(iv) concerns the controller using the model in an AI system. That controller may not develop the model or control its filters. Effective monitoring therefore depends both on what it can observe and on what action a detection can produce. A generic instruction to monitor does not supply either.

Proposed addition to paragraph 74(iv)

> Where monitoring depends on information or action from another entity, the controller using the model should establish a workable information and response arrangement. Information supplied as part of a relied-upon safeguard should identify the relevant model version or configuration, the scope and material limitations of the protection, how detected disclosures are reported and acted upon, and how changes affecting those conditions are communicated. The responsibilities of the entities concerned follow their GDPR roles, not merely their commercial descriptions as developer or user.
>
> The controller should address any missing information or unavailable response that prevents the measure from performing its intended function. An adequate alternative measure or a restriction of the affected processing may be necessary; the information gap should not be treated as evidence that monitoring is effective or that responsibility has transferred elsewhere. Monitoring records should be limited to what is necessary, with appropriate access and retention controls, rather than routinely accumulating another corpus of sensitive outputs.

4. Assess a claimed limitation at the stage where it matters

Suggested placement: after paragraph 75; read with paragraphs 73–74 and 77.

Paragraph 75 asks for evidence of the conditions and measures described in paragraphs 73–74. A statement that a supplier cannot act leaves a decisive question unanswered: is the obstacle a limit of the available techniques, an information gap or an action that has not been organised? The explanation should identify the processing stage as well as the cause. The assessment remains case-specific; organisational factors are not categorically excluded from it.

Paragraph 73(c) concerns assessing and preventing collection. Two separate questions can otherwise be conflated: why relevant material could not be detected before collection, and why material already identified was not removed or excluded from further use. A missing task owner, unavailable language coverage, discarded source reference or missing supplier response may explain the second without establishing the first. Conversely, escalation cannot supply a technique that does not exist.

Proposed wording

> The controller should identify the processing stage, action and evidence to which a claimed limitation relates. It should distinguish limitations in the available techniques from limitations in information, expertise, task allocation, permissions or the ability to obtain action from another entity. A supplier's inability to act or an unassigned task does not, without further examination, establish the difficulty referred to in paragraph 73(c). Equally, the existence of an escalation route does not establish that the required technical action is feasible.
>
> Where the cause or extent of a limitation is not established, it should be recorded as unresolved rather than assigned an unsupported explanation. Difficulty in predicting or preventing collection should not be treated as an explanation for inaction once relevant material has been identified. The controller should separately assess the available measures, the cooperation needed to perform them and the implications for further processing.

This is a proposal about evidentiary discipline if the draft retains its analogy with *GC and Others*. The judgment concerns search-engine processing; the response does not claim that it creates a general exemption for training generative models. Nor does it propose a new Article 9 exception [5].

5. Test the operating conditions on which safeguards depend

Suggested placement: addition to paragraph 76, followed by a practical illustration.

A filter can retain its settings while a supplier changes its language coverage, volumes exceed review capacity or a permission change removes the ability to halt training. Verification confined to configuration would miss these changes. Where human review contributes to the measure, the relevant question is what the person can understand and cause to happen, not whether a reviewer remains named.

Proposed wording

> Verification should consider the conditions on which a measure's effectiveness depends, including source and language coverage, volume, filter configuration, provider responsibilities, relevant expertise, access and the route to corrective action. Where human review is relied upon, it should examine whether reviewers have the relevant information and competence, sufficient capacity for their assigned work, and the authority or practical route needed for timely intervention. A change to those conditions should be assessed for its effect on all processing relying on the measure.

A controlled exercise could trace a synthetic removal instruction from receipt to exclusion from a test training input. It should include a transformed copy and an unavailable usual contact. No real data-subject request needs to be fabricated, and no real person's rights should be affected. The exercise can expose a broken handoff; it cannot establish the filter's detection performance, the absence of special-category data or legal compliance. Those require their own evidence.

Existing records and processes should be used where sufficient. An organisation may combine roles, provided the distinct decisions and tasks remain clear. A data protection officer can advise and monitor within the statutory role; allocating the operational response should not silently make the DPO the business decision-maker. Evidence references should be preferred to unnecessary copies of personal data.

Illustrative example for the Guidelines

A company obtains a multilingual scraped dataset for fine-tuning. After delivery, its supplier identifies health disclosures in a batch that passed the supplier's filter. The company has already transformed the batch and staged it; training has not started. Assume that no Article 9(2) condition supports retaining or using those disclosures.

The response needs to reach the transformed training input, not stop at the supplier's record. A shared batch reference connects the finding to the delivered material; a designated function at the company can prevent its admission to training and arrange removal; the completion record states which copies were addressed and what remains unresolved. Prevention of further use is not a basis for retaining the disclosures indefinitely. If the supplier cannot identify the affected delivered material, that information limitation may require a wider exclusion while the necessary action is determined; it is not, by itself, proof of technical impossibility.

The filtering assessment and the legitimate-interest assessment are reconsidered where the finding affects their premises. Any implications for earlier training or an already trained model are assessed separately. A repaired filter and a cleared staging area do not, by themselves, settle those questions. This is a hypothetical example, not a reported client case.

Closing recommendation

The final Guidelines should connect each relied-upon safeguard to the material it covers, the conditions under which it operates and the actors able to give it effect. This would make evidence of protection more informative without confusing supplier delivery with justified reliance, technical limitation with unfinished implementation, or technical correction with the decisions still required.

About the contributor

Maja Kurek, AIGP, CIPM, CIPP/E, is Founder and Principal of RuleBridge Advisory, an independent AI governance practice based in Zagreb, Croatia, working on the operator and deployer side of AI arrangements: organisational AI governance, human oversight and decision accountability, and the governance of external AI relationships with vendors, platforms, models and external contributor networks. Her operational background spans close to a decade across Google, Photomath and Uber, including full-lifecycle responsibility for the human layer of an AI-powered learning product, governing a global network of more than 10,000 external experts (admission, access, cross-border payment, data protection, offboarding), data and access controls, regulatory adaptation programmes for a platform operating under changing national law, and governance-sensitive provider transitions. This contribution is made through RuleBridge Advisory, which provides advisory services in related areas. It does not represent any former employer or client.

Notes and sources

[1] European Data Protection Board, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted for public consultation on 7 July 2026; in particular paragraphs 14, 16–20, 32–33, 64–66 and 72–77.
https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf

[2] CNIL, AI system development: CNIL's recommendations to comply with the GDPR, especially Steps 3(bis), 4, 8 and 9.
https://www.cnil.fr/en/ai-system-development-cnils-recommendations-to-comply-gdpr

[3] Regulation (EU) 2016/679 (GDPR), Articles 5, 6, 9, 12, 17–19, 21, 24–26, 28 and 38–39. See also EDPB Opinion 28/2024, paragraphs 102 and 106, for the distinction between additional advance opt-out safeguards and statutory objection rights.
https://eur-lex.europa.eu/eli/reg/2016/679/oj
https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf

[4] Kurek, M. (2026), *Governing AI After Deployment: Field Evidence on Model Changes, Platform Configuration and Organisational Authority*, SSRN working paper, sections 3–5. DOI: 10.2139/ssrn.7329378. Participant-based evidence from one arrangement, with limited screen observations and factual confirmation; not a study of scraping or an evaluation of these recommendations.
https://doi.org/10.2139/ssrn.7329378

[5] EDPB Guidelines 03/2026, paragraphs 70–77, discussing *GC and Others*, C-136/17, judgment of 24 September 2019, especially paragraphs 45 and 47–48. The proposal concerns the draft's evidentiary approach, not an extension of the judgment by this contributor.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62017CJ0136