COMMENTS ON EDPB GUIDELINES 03/2026

Operational Evidence and Control Implementation for Web Scraping in Generative AI Development

Submitted by	Solutions by Jewel (SBJ)
Submitter	Jewel Darlington, Chief AI Governance Officer & Founder
Organization	Solutions by Jewel, a global AI governance advisory firm
Contact	governance@solutionsbyjewel.com
Website	solutionsbyjewel.com

Consultation	EDPB Guidelines 03/2026 on web scraping in the context of generative AI
Deadline	30 October 2026, 23:59 CET
Date	10 October 2026

Purpose and scope

Solutions by Jewel (SBJ) provides enterprise AI-governance architecture and operational accountability analysis. This submission addresses how controllers and their service providers can make selected safeguards in Guidelines 03/2026 implementable, observable and reviewable across data acquisition and model development.

SBJ welcomes the Board's treatment of collection criteria, data minimisation, transparency, third-party relationships, safeguards and continuous verification. Each recommendation below seeks a limited, illustrative implementation example that operates within the Board's existing interpretation of the GDPR. Each comment closes with suggested wording the Board may adapt.

Summary of requested clarifications

No.	Issue	Guideline ¶¶
01	Dynamic enforcement of collection scope	8, 10–11, 36–38
02	Third-party provenance and processing responsibilities	16–20, 32
03	Verification of minimisation and exclusion safeguards	35–38, 65–66
04	Change-triggered reassessment of existing decisions	52–53, 64–66
05	Evidence of safeguards and effective intervention	72–77

Primary source: EDPB Guidelines 03/2026, version 1.0 (adopted 7 July 2026) [1]. Secondary authority: EDPB Opinion 28/2024 (adopted 17 December 2024) [2]. Full URLs appear in the reference list.

Interpretive conventions applying throughout

The following conventions apply to every comment, so they are stated once here.

Legal determinations. Lawful basis, Article 9 conditions, Article 14 exceptions and controller, processor or joint-controller classification remain fact-specific legal determinations for the controller and the competent supervisory authority. The suggested records support those determinations by documenting the facts on which they rest.

Proportionality. Suggested records, decision gates and verification steps scale with the context, risk and technical feasibility of the processing. Where a comment lists illustrative evidence, the list is non-exhaustive and every item is optional.

Terminology. “Authorisation” refers to internal operational approval and control scope. Lawful basis for processing is assessed separately under Article 6 GDPR.

Research evidence. References [9]–[12] are peer-reviewed technical studies. They document implementation challenges observed in specific datasets and systems; their findings are cited for that operational relevance.

Contributor resources. SBJ’s published AI Governance Architecture (V1.0 and V2.0) and Decision Record V2.0 [5]–[7] are original contributor methodologies, cited as worked examples of governance design. Their status is that of contributor material; each comment refers to them in a single sentence, and the detailed control mappings appear in Annex A for readers who wish to consult them.

Comment 01. Dynamic enforcement of collection scope

Target paragraphs: 8, 10–11 and 36–38 (printed pp. 6–7 and 12 of Guidelines 03/2026).

Existing guidance

Paragraph 8 describes how a crawler may continuously discover new URLs, increasing the risk that the controller has limited knowledge of the data collected. Paragraph 11 identifies criteria and source exclusions at the selection stage. Paragraph 37 recommends precise collection criteria, mapping and exclusion of relevant sites or content.

Implementation gap

An approved collection plan can drift from the actual crawl as links, source characteristics, site access conditions or crawler configurations change. A pre-collection assessment records the intended boundary; showing that the boundary held during each collection run calls for run-level evidence. Without it, the controller loses the ability to separate data collected within the reviewed criteria from material source expansions or missed exclusions, which weakens both the demonstration of collection scope and the basis for later corrections.

Requested clarification

A short implementation example describing a versioned acquisition-control record. The record would link each material collection run to the approved purpose, source-selection criteria, applicable restrictions and exclusions, crawler configuration, detected boundary changes and the accountable disposition. The example could explain that a material scope change triggers review or restriction before affected data progresses into training, where the controller’s documented controls require it. This builds on the Board’s existing exclusion criteria by evidencing that they remained effective as the crawl evolved; machine-readable signals such as robots.txt or ai.txt would be recorded as one input to that assessment.

Illustrative evidence

Run identifier; applicable source policy and version; domain or source-class decision; time a restriction was observed; filter configuration and test result; exception or deviation; person or function accountable for review; and release, quarantine, re-collection or exclusion disposition. Linked manifests can carry this evidence, so raw content need only be retained where the controller otherwise has a basis to keep it.

Suggested wording

> *Insert after paragraph 37:* “For example, a controller could maintain a versioned record linking each material collection run to the approved collection criteria, applicable exclusions and crawler configuration, together with any detected deviation and the resulting decision on whether the affected data proceeds to model training.”

Supporting evidence

Research on Common Crawl documents filtering and documentation limitations in widely used web-crawl data, supporting verification of the controls actually applied at collection time (Baack, 2024) [10]. SBJ’s published architecture applies the same principle by tying each consequential execution to its run and configuration version (Annex A).

Comment 02. Third-party provenance and processing responsibilities

Target paragraphs: 16–20 and 32 (printed pp. 8–9 and 11 of Guidelines 03/2026).

Existing guidance

Paragraphs 16–20 distinguish factual controller, processor and joint-controller circumstances, including documented instructions in contracted scraping and separate processing by an independent dataset provider. Paragraph 32 recommends clearer information about the source controller and collection conditions for already-scraped datasets.

Implementation gap

The guideline identifies the possible parties and their responsibilities. Practical detail would help on preserving the evidence chain that links a supplied dataset to its source conditions, any contractual instructions, its transformations and its use in a specific model-development version. Where that chain is incomplete, an AI developer may struggle to reconstruct what was obtained, from which supplier, under what documented conditions and where it was incorporated, which reduces the reliability of due diligence, source-specific corrective action and downstream accountability.

Requested clarification

An illustrative provenance record connecting: acquisition-party identity and factual role assessment; collection period and source categories; documented source restrictions and supplier representations; documented instructions where the relationship is processor-based; diligence limitations; dataset identifiers and material transformations; and a training-run or model-version reference. The record would separate evidence received from suppliers from facts the developer has independently verified. The example would apply across supplier relationships of every kind, with record granularity set by the controller’s risk assessment.

Illustrative evidence

Supplier and dataset manifest; collection date range; source-class summary; asserted collection conditions with verification status; documented processing instructions where applicable; accepted limitations; transformation and version lineage; and the reviewer’s use-or-restrict decision.

Suggested wording

> *Insert after paragraph 32:* “Where personal data is obtained from a third-party dataset provider, the developer may document the provider’s identity and role, the collection conditions it has represented, the extent to which those representations were independently verified, and the training runs or model versions in which the dataset was used.”

Supporting evidence

An audit of 1,858 text datasets found substantial gaps and inconsistencies in licensing and attribution metadata, illustrating the value of checking inherited supplier descriptions against original-source provenance (Longpre et al., 2024) [9]. In the context of assessing whether an AI model is anonymous, Opinion 28/2024, paragraph 58(c)–(f), identifies source-specific documentation, third-party measures, lifecycle records and testing evidence as relevant [2]; SBJ suggests the same documentation discipline transfers usefully to scraping provenance. SBJ’s lineage and shared-responsibility controls are mapped in Annex A.

Comment 03. Verification of minimisation and exclusion safeguards

Target paragraphs: 35–38 and 65–66 (printed pp. 12 and 18–19 of Guidelines 03/2026).

Existing guidance

Paragraphs 35–38 address collection necessity, source mapping, filtering, exclusions and later transformations. Paragraphs 65–66 distinguish GDPR-required measures from additional mitigating safeguards that may be relevant to legitimate-interest balancing.

Implementation gap

Four distinct facts bear on whether a filter or exclusion rule works: its configuration, a test of its operation, the failures or coverage limits that test reveals, and the resulting decision about affected data. Where only the first is documented, unnecessary or sensitive personal data can reach cleaned training datasets with no clear basis for judging the control’s effectiveness or remediating its failures.

Requested clarification

A compact example of a safeguard-verification record, focused on high-impact source exclusions and data filters. The record would identify the control objective, the data and source classes assessed, the test method, sampling or coverage limitations, observed false negatives and exceptions where measurable, remediation, the retest result and the accountable determination on whether the affected data can be used. Verification intensity would be risk-based and technically feasible, and the example would remain illustrative of good practice.

Illustrative evidence

Control version and scope; evaluation sample and its rationale; before-and-after results; known detection limitations; unresolved exceptions; reviewer and date; corrective action; and subsequent evidence of effect. Results would record demonstrated performance and untested assumptions separately.

Suggested wording

> *Insert after paragraph 38:* “Controllers may document how the effectiveness of key filtering and exclusion measures was tested, including the method used, known limitations, residual failures observed and the decision taken on the affected data.”

Supporting evidence

A study of a large web-scraped machine-learning dataset identified personal information that remained after sanitisation (Hong et al., 2026) [11]; its findings concern the dataset examined and support measuring filter performance case by case. In its anonymity-assessment context, Opinion 28/2024, paragraph 58(e), lists documentation of model testing: by whom, when, how, to what extent and with what result [2]. Applying that evidentiary discipline to collection-stage controls is SBJ's proposed analogy. SBJ's evidence-sufficiency concept is mapped in Annex A.

Comment 04. Change-triggered reassessment of existing decisions

Target paragraphs: 52–53 and 64–66 (printed pp. 15 and 17–19 of Guidelines 03/2026).

Existing guidance

Paragraphs 52–53 address necessity and the balancing of relevant interests and rights. Paragraph 64 notes that the impact of generative AI models may change with model capability and identifies contextual factors relevant to reasonable expectations. Paragraphs 65–66 address case-specific mitigating measures.

Implementation gap

An assessment or control decision can become operationally stale when material circumstances change after approval: a new category of source, expanded scraping reach, changed access restrictions, new third-party data dependencies, an altered model purpose or newly observed safeguard failures. Collection or training can then continue on historical evidence that no longer describes the present processing or the effectiveness of its safeguards.

Requested clarification

Illustrative material-change triggers, together with a decision record connecting the changed fact to the affected processing purpose, data source or dataset, existing assessment, reassessment, responsible decision-maker and resulting outcome: continuation, narrowing, suspension or remediation. The guidance could separate events calling for documented operational review from questions calling for qualified legal assessment. The example would be reserved for material changes, with the controller determining materiality and the proportionate response.

Illustrative evidence

Change signal and date; affected source and dataset or model version; earlier assessment reference; nature and materiality of the change; assessment requested and completed; accountable reviewer; interim containment where warranted; and recorded outcome with next review condition.

Suggested wording

> *Insert after paragraph 64:* "Controllers may identify in advance the changes they consider material, such as new source categories, expanded collection reach or observed safeguard failures, and document how each such change was reassessed and what decision followed."

Supporting evidence

The guideline itself recognises in paragraph 64 that context and model capability evolve; a worked example would show how that observation carries into ongoing governance. SBJ's published material-change profile is mapped in Annex A.

Comment 05. Evidence of safeguards and effective intervention

Target paragraphs: 72–77, especially paragraph 74(i)–(iv) and paragraphs 75–76 (printed pp. 20–22 of Guidelines 03/2026).

Existing guidance

Paragraphs 72–74 discuss case-specific considerations for incidental and residual processing of special categories of personal data and describe measures before collection, after collection, during model development and after deployment. Paragraphs 75–77 discuss demonstrability, verification and state-of-the-art safeguards.

Implementation gap

The guideline sets out the relevant safeguards. It could illustrate more fully how a detected failure is traced to affected datasets or model versions, how the responsible function records its intervention decision, and how the effectiveness of remediation is verified. A controller that detects data which should not have entered a dataset, or a problematic model output, needs that chain to demonstrate that remediation covered the affected processing stages and that any release or continuation decision was justified.

Requested clarification

One non-prescriptive, end-to-end example: (1) detect and classify the finding; (2) identify affected sources, processing stages and model versions where feasible; (3) restrict, delete or otherwise remediate within the controller's responsibilities; (4) test and record the outcome, including residual limitations; and (5) document the accountable continuation, restriction or cessation decision. The example would treat removal from a dataset, model-output controls and verified removal from model parameters as three separate measures, each with its own evidence. This comment concerns evidence of effectiveness; the Article 9 analysis in paragraphs 72–74 remains specific to each case.

Illustrative evidence

Finding and discovery time; affected asset and version, with traceability limits; intervention owner; immediate containment; selected measure and rationale; privacy-attack or output retest and its limitations; unresolved residual effects; and signed-off decision with a next verification date where warranted.

Suggested wording

> *Insert after paragraph 76:* "Where a safeguard failure is detected, controllers may document the affected datasets and model versions, the remedial measure applied, how its effect was tested, and the resulting decision on continued processing. Removal from a training dataset, output-level controls and removal from model parameters are distinct measures and are evidenced separately."

Supporting evidence

Security research shows that filtering, preprocessing and output controls can themselves create privacy side channels in some architectures (Debenedetti et al., 2024) [12], so evaluating a remedial measure calls for examining the assembled system against its threat model. SBJ's intervention and recovery controls are mapped in Annex A.

Cross-cutting clarification

Together, the five comments propose a proportionate record of three linked events: (i) the approved or assessed processing scope; (ii) what happened during data acquisition, transformation and model development; and (iii) the response to changes or observed control failures. Keeping these events distinguishable supports oversight through targeted records, without exhaustive raw-data logging or the collection of additional personal data for documentation purposes.

Each kind of claim carries its own evidence. A training-run manifest or dataset receipt shows that a dataset was supplied to a pipeline; model-level causal influence, complete erasure and effective prevention of disclosure are each evidenced separately. Sensitive raw personal data stays within the controller's systems throughout.

Scaling for smaller controllers. The suggested records scale down. A small developer running a limited crawl could meet every example in this submission with a single manifest per collection campaign, recording scope, exclusions applied, tests run, exceptions found and the decision taken.

The examples are technology-neutral. Evidence may sit in source inventories, dataset manifests, logs, change-control systems, assessment registers or linked records, at a level of detail suited to the controller's circumstances, risks and technical feasibility. The Board's legal standards remain as set out in the guideline.

Annex A — Mapping to SBJ contributor resources

This annex is provided for readers who wish to see how the comments draw on SBJ's published methodologies [5]–[7]. Identifiers are shown with a hyphen for consistency (the Architecture V2.0 source labels VE-01 and VE-02 as VE01 and VE02).

Comment	SBJ resource and control
01	Architecture V2.0 §5, AC-01 (reassessment when authorisation-supporting conditions change) and AC-04 (linking consequential execution to its execution instance and version or configuration); §6 (reconstructing a material execution path) [5]. Decision Record V2.0: separation of governance disposition from operational records [6].
02	Architecture V2.0 §6, VE-01 (transformation and decision lineage) and VE-02 (availability of data to a training environment, evidenced separately from causal influence on model computation); §10 and AC-07 (shared-responsibility deployments) [5]. Decision Record V2.0 conditional Profile G [6].
03	Architecture V2.0 §5, AC-05 (evidence existence and evidence sufficiency for the decision supported) [5]. Decision Record V2.0 evidence-sufficiency determination and stable references to authoritative records [6].
04	Architecture V2.0 §5, AC-01 [5]. Decision Record V2.0 conditional Profile A (material change), with Profile G where reassessment evidence depends on external suppliers [6].
05	Architecture V1.0 adverse-evidence preservation and intervention [7]. Architecture V2.0 §5, AC-05, and its observation that ending execution may leave committed downstream effects in place [5]. Decision Record V2.0 conditional Profile F (incident, corrective action and recovery) [6].

References

[1] EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted 7 July 2026. https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf

[2] 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; esp. paragraphs 56–58.

https://www.edpb.europa.eu/system/files/documents/2024-12/edpb_opinion_202428_ai-models_en.pdf

[3] EDPB, Guidelines 07/2020 on the concepts of controller and processor in the GDPR. https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en

[4] Regulation (EU) 2016/679 (GDPR); notably Articles 5, 6, 9, 24, 25, 28, 30 and 35. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

[5] Solutions by Jewel, SBJ AI Governance Architecture V2.0, particularly sections 5, 6 and 10 (original contributor architecture). https://solutionsbyjewel.com/ai-governance/architecture/

[6] Solutions by Jewel, SBJ AI Governance Decision Record V2.0, methodology LOCATE → ASSIGN → RECORD → SCALE, core evidence references and sufficiency determination, and relevant conditional profiles (original contributor decision-documentation method). https://solutionsbyjewel.com/ai-governance/decision-record/

[7] Solutions by Jewel, SBJ AI Governance Architecture V1.0, particularly evidence, accountability, continuous control and adverse-evidence preservation (original contributor methodology, subject to its stated scope). https://solutionsbyjewel.com/sbj-ai-governance-architecture-v1

[8] EDPB, consultation portal and publication/submission instructions. https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en

[9] Longpre, S., Mahari, R., Chen, A. et al. (2024), A large-scale audit of dataset licensing and attribution in AI, *Nature Machine Intelligence* 6, 975–987. https://doi.org/10.1038/s42256-024-00878-8

[10] Baack, S. (2024), A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl, ACM FAccT 2024. https://doi.org/10.1145/3630106.3659033

[11] Hong, R., Hutson, J., Agnew, W. et al. (2026), A Common Pool of Privacy Problems: Legal and Technical Lessons from a Large-Scale Web-Scraped Machine Learning Dataset, ACM Symposium on Computer Science and Law. https://doi.org/10.1145/3788646.3789521

[12] Debenedetti, E., Severi, G., Carlini, N. et al. (2024), Privacy Side Channels in Machine Learning Systems, USENIX Security Symposium 2024. https://www.usenix.org/conference/usenixsecurity24/presentation/debenedetti

Contributor

Solutions by Jewel (SBJ) · Enterprise AI Governance and Operational Accountability
https://solutionsbyjewel.com