TERLOUW.LEGAL / PUBLIC CONSULTATION

GUIDELINES 03/2026

Web scraping for generative AI

Technical comments on evidence, provenance and auditable control

17 September 2026 | Netherlands

CORE CONTRIBUTION

A scraping control that cannot later be linked to the source, time, dataset and decision it governed is difficult to audit and difficult to defend.

terlouw.legal is a Dutch legal services and legal technology organisation. Our work focuses on legal control in situations where technical facts, evidence, governance and legal rules must remain connected over time. The draft Guidelines already identify many substantive safeguards. This contribution addresses a narrower implementation question: what evidence should exist so that those safeguards can later be verified?

SCOPE OF THIS SUBMISSION

This contribution focuses on operational accountability, evidence quality and decision traceability. It does not propose a new technical standard, statistical threshold or sector specific rule. The aim is to test whether the draft can be implemented in a way that remains auditable after the relevant event.

Organisation: terlouw.legal | Amsterdam, Netherlands | https://terlouw.legal

Contact details are provided through the EDPB submission form and are not repeated in this public PDF.

Executive summary

The draft connects transparency, data minimisation, source restrictions, legitimate interest, rights handling and model safeguards. The operational risk is that these controls are implemented in separate systems and cannot later be reconstructed as one decision trail. That becomes especially important when source settings change, a dataset is re-used, an objection arrives after training, or the organisation must explain what a crawler encountered months earlier.

PROPOSED IMPLEMENTATION LENS

Treat provenance as control evidence, not only as descriptive metadata. For each material dataset or collection run, the organisation should be able to reconstruct: Fact, Source, Rule, Decision and Action.

Six points for clarification

- Source transparency should be supported by contemporaneous evidence of what was collected, when and under which crawler configuration.
- Machine-readable opposition signals should be captured at collection time because the signal may later change. Their absence is not consent.
- A purchaser of an already scraped dataset needs a practical evidence handoff, not only a contractual assurance from the supplier.
- Rights handling needs lineage from an identifiable source reference to affected dataset stages and, where relevant, model stages.
- A legitimate interest assessment should be tied to a defined purpose, dataset version and control set so that later changes trigger review.
- Audit evidence itself should be designed to minimise personal data and avoid recreating the very dataset that a controller seeks to limit.

01 Make source provenance a contemporaneous record

Draft reference: paras. 30-32, 37-38 and 66(e)

Observation. The draft asks for precise source information, crawler characteristics, searchable source information where possible, data mapping and an updated list of scraped websites.

Implementation issue. A later list of domains does not necessarily show what happened during a specific collection event. Pages, access controls, robots.txt, ai.txt, APIs and site structure can change. A post hoc screenshot may therefore be weak evidence of the factual conditions at the time of collection.

Clarification requested. Please clarify whether accountability can be supported by a contemporaneous scrape manifest that records the collection period, crawler identity and version, source or source class, relevant machine-readable signals, inclusion or exclusion rule version and the dataset partition receiving the data. The manifest can store hashes or metadata where retaining content would conflict with minimisation.

> **PRACTICE EXAMPLE**
>
> A site permits crawling in September but changes its robots.txt in November. A complaint in December cannot be assessed reliably from the November configuration alone. A timestamped collection record can show which signal the crawler actually encountered.

02 Separate opposition signals from consent, and preserve the signal state

Draft reference: paras. 37, 45 and 64

Observation. The draft correctly distinguishes technical opposition to scraping from GDPR consent and notes that the absence of a robots.txt file does not amount to consent.

Implementation issue. In practice, teams may treat robots.txt, ai.txt, CAPTCHA, login requirements, terms and platform settings as a single binary field called permission. That can blur different legal meanings and can hide whether a control was a source restriction, an expectation factor, a technical barrier or a copyright signal.

Clarification requested. Please clarify that operational records should preserve the type of signal and its role in the assessment. A useful record distinguishes at least access status, anti-bot control, machine-readable scraping signal, relevant source terms and the controller decision. The evidence should record the signal observed at the relevant time, not merely the current state.

03 Create an evidence handoff for re-used scraped datasets

Draft reference: paras. 18-19, 32 and 43

Observation. The draft recognises separate responsibility where an AI developer obtains a dataset already scraped by another organisation and requires the developer to identify its own lawful basis for re-use.

Implementation issue. A downstream controller may be unable to perform a meaningful assessment if it receives only the files and a warranty that the data was lawfully collected. The most useful facts may sit with the original scraper: source classes, collection dates, exclusions, crawler configuration and rights signals.

Clarification requested. Please clarify the minimum provenance information that a downstream controller should seek and assess before re-use. A proportionate evidence package could identify dataset version, collection period, source classes, material exclusions, legal basis asserted by the collector, rights mechanism, known special category controls and any unresolved limitations. This would not transfer legal responsibility; it would make the downstream assessment evidence-based.

> **PRACTICE EXAMPLE**
>
> A dataset vendor says that all data came from public sources but cannot identify source categories, collection dates or exclusion rules. The buyer can document the supplier statement, but cannot test reasonable expectations, source restrictions or data minimisation in a meaningful way.

04 Connect rights handling to data and model lineage

Draft reference: paras. 60 and 66(f)-(i), and paras. 74-75

Observation. The draft recognises that exercising rights can be difficult after large-scale scraping, that deletion from a trained model may be technically difficult, and that model safeguards, monitoring and future unlearning can become relevant.

Implementation issue. A rights process is only operational if the controller can determine what an objection or erasure request maps to. The same source material may exist in raw collections, cleaned datasets, deduplicated training sets, fine-tuning data, evaluation corpora, checkpoints and released model versions.

Clarification requested. Please clarify that a controller should maintain proportionate lineage evidence showing how a source reference or data subject identifier can be mapped to affected processing stages where identification is possible. The record should also distinguish actions that remove data, actions that suppress future collection, and actions that mitigate model output when deletion or unlearning is not yet technically available.

> **PRACTICE EXAMPLE**
>
> This does not require keeping extra identifiers solely to facilitate rights. It requires designing the pipeline so that, where the controller can identify the relevant data, the effect of a request can be traced and evidenced.

05 Bind the legitimate interest assessment to a versioned processing reality

Draft reference: paras. 43-66

Observation. The legitimate interest assessment is case specific and depends on the interest pursued, necessity, source context, data categories, reasonable expectations, consequences and mitigating measures.

Implementation issue. Generative AI development changes quickly. A balancing test prepared for one purpose and control set can become stale when the collection scope expands, new source classes are added, the model capability changes, or safeguards are removed or replaced.

Clarification requested. Please clarify that the assessment is best treated as a versioned decision record tied to a defined purpose, dataset or collection scope and control set. Material changes should create a review event. This makes it possible to explain which balancing test governed which collection and model development activity without suggesting that a single assessment is valid indefinitely.

06 Design accountability evidence so it does not recreate the dataset

Draft reference: paras. 30-38 and Article 5(2) GDPR context

Observation. The draft calls for accountability, source transparency, data mapping and evidence of safeguards. At the same time, minimisation and storage limitation require avoiding unnecessary personal data.

Implementation issue. An overly literal audit design could store copies of pages, personal identifiers or detailed raw logs merely to prove that a control ran. That can expand the personal data footprint and create a second retention problem.

Clarification requested. Please clarify that evidence of compliance can often be satisfied through minimised control metadata. Examples include source-domain records, hashes, rule identifiers, timestamps, aggregate filter results, dataset partition IDs and deletion receipts. Raw personal data should be retained for audit only where necessary for a defined purpose and retention period.

ANNEX / SCRAPE EVIDENCE RECORD

The following is not proposed as a mandatory template. It illustrates the level of traceability that can connect the draft Guidelines to an auditable implementation without storing unnecessary scraped content.

FIELD	EVIDENCE TO RETAIN
Processing purpose version	Identifier and short statement linking the collection run to the specific development purpose and lawful basis assessment.
Collection event	Start and end time, crawler or scraper ID, software version, operator or service, and run identifier.
Source evidence	Domain or source class, access status, URL where proportionate, collection date, machine-readable opposition signals observed, and evidence hash or snapshot reference.
Decision rules	Inclusion and exclusion rule version, excluded source classes, personal data filters, special category controls, and any exceptions.
Dataset lineage	Raw partition ID, cleaned dataset ID, training or fine-tuning set ID, model or checkpoint versions materially linked to the dataset.
Legal decision	Lawful basis record, legitimate interest assessment version if used, necessity analysis, reasonable expectation factors and mitigation set.
Rights and objections	Opt-out or objection reference, match method, suppression action, deletion or anonymisation action, downstream propagation status and evidence of completion.
Review event	Material change trigger, reviewer, date, changed fact, resulting decision and action.

WHY THIS MATTERS

The core question is not whether an organisation has a policy. It is whether, months later, it can show which fact was known, which source and rule applied, which decision followed, and what action was actually executed.

Closing observations

The draft already contains substantial guidance on substantive safeguards. The remaining implementation challenge is evidentiary. Large-scale AI data pipelines create distance between the original web source, the collection event, the legal decision and the final model. That distance increases the risk that a controller can describe its controls in general terms but cannot prove how they applied to the dataset or event under review.

terlouw.legal therefore invites clarification on whether the Board considers a proportionate, versioned evidence record consistent with the accountability principle. The record need not be a prescribed format. Its value lies in preserving the connections between facts, sources, legal rules, decisions and executed actions.

THREE QUESTIONS FOR THE FINAL GUIDELINES

1. What minimum source and collection evidence should a controller retain to substantiate its transparency, minimisation and expectation assessments?
2. What minimum provenance should a downstream AI developer obtain when re-using a previously scraped dataset?
3. How should controllers evidence the propagation of objections, deletion or mitigation across dataset and model stages without retaining unnecessary identifiers?

Source basis

This submission comments on the public consultation version of the EDPB draft. Paragraph references in the body refer to that version.

- **EDPB Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted for public consultation on 7 July 2026:** https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf
- **EDPB public consultation page:** https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en
- **EDPB Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models:** https://www.edpb.europa.eu/our-work-tools/our-documents/opinion-board-art-64/opinion-282024-certain-data-protection-aspects_en
- **EDPB Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR:** https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-12024-processing-personal-data-based-article-61f_en

PUBLICATION NOTE

The EDPB consultation page states that submitted comments are published after screening for unauthorised submissions. The PDF therefore contains no personal signature and no personal contact details. The organisation name is intentionally visible.

FACT. SOURCE. RULE. DECISION. ACTION.