

Article 9 Web-Scraping Governance Checklist

A lifecycle map for incidental and residual special-category data in generative-AI development

Use this checklist to test whether an “incidental” Article 9 classification is supported by the system’s purpose, source controls, technical performance and retained evidence. It is designed for design reviews, dataset acquisition, model release, incident response and supplier oversight.

Legal baseline

Intentional processing	A lawful basis under Article 6 and an applicable derogation under Article 9(2) are required.
Incidental and residual processing	The EDPB describes a narrow, case-by-case pathway tied to relevant search-engine similarities and the controller’s responsibilities, powers and capabilities.
No general safe harbour	The Guidelines are consultation-stage guidance. Other GDPR duties still apply, and the position must be tested against binding CJEU case law.

The EDPB’s four-condition test

The controller should document a case-by-case assessment of all four conditions. The examples below are illustrative only; they do not create a safe harbour.

1	Relevant similarity to a search engine The processing has relevant similarities to search-engine activity. <i>Example: a general-purpose crawler systematically gathers publicly accessible webpages without targeting sources because they contain special-category data</i>
2	Incidental and residual only The special-category processing is unwanted residue, not intentional. <i>Example: a medical diagnosis happens to appear in a general discussion forum included in a broad training corpus.</i>
3	Difficult or impossible to identify in advance The controller cannot reliably determine before collection whether, or to what extent, Article 9 data will be included. <i>Example: political opinion or sexual orientation is revealed only by the meaning and context of free text, not by a detectable field or URL.</i>
4	Measures match responsibilities, powers and capabilities The controller implements effective measures to prevent collection and dissemination. <i>Example: exclude structurally sensitive sources, filter likely Article 9 content, delete detected material and prevent it resurfacing in outputs.</i>

Source: EDPB Guidelines 03/2026, paras 72-74 (especially para 73).

Lifecycle governance checklist

For each stage, record the decision owner, status, evidence location and date of the most recent review in the organization's system of record.

Lifecycle stage	What the organization must do	Evidence to retain
Scoping	Define the purpose and why special-category data are unnecessary. Identify the point at which their presence would cease to be incidental. Set escalation and redesign thresholds.	Purpose statement; Article 9 assessment; excluded functions and data; escalation thresholds; legal and governance sign-off.
Sourcing and collection	Map source types. Exclude structurally sensitive, minors-focused and restricted environments. Assess whether content is genuinely freely accessible. Respect clear anti-scraping signals. Test collection filters.	Source register; exclusion rules; accessibility assessment; crawler configuration; robots.txt/CAPTCHA decisions; filter test results; change log.
Post-collection	Detect and quarantine suspected Article 9 data. Delete it immediately after collection or as soon as identified. Test false negatives. Trace derived copies, indexes and relevant backups.	Detection rules and model cards; validation results; quarantine and deletion logs; data lineage; derived-copy records; data-subject request record.
Training and evaluation	Prevent extraction from the model. Test memorisation, regurgitation, membership inference, inversion and other privacy attacks. Validate output filters against realistic prompts and attacks.	Evaluation protocol; attack prompts and datasets; thresholds; results; red-team reports; remediation decisions; release criteria and approval.
Deployment	Continuously monitor outputs and connect alerts to incident and change control. Contain sensitive outputs, investigate causes, reinforce filters or restrict prompts/functions, and assess retraining or effective unlearning.	Monitoring design; privacy-minimised logs; alerts; incident reports; containment actions; filter/model versions; retraining or unlearning assessment.
Supply chain	Map controller, joint-controller and processor roles. Require evidence and cooperation from dataset vendors, model developers and deployers. Reassess the incidental classification for each actor's own processing.	Responsibility allocation; contract clauses; vendor source/filter evidence; audit rights; incident cooperation; deletion/retraining support; testing commitments.

Operational synthesis of EDPB Guidelines 03/2026, paras 74-77, supplemented by CNIL's January 2026 web-scraping measures.

Cross-cutting review. Regularly verify and assess whether the measures remain relevant and effective throughout development and use. Monitor state-of-the-art techniques for limiting the collection and other processing of special-category data, and update controls when more effective measures become available. Retain review records, test results and resulting change decisions.

When “incidental” stops being incidental

WARNING

Purpose. The use case, expected function or evaluation objective requires special-category data.

Sources. Collection deliberately targets health forums, political groups, religious communities or other structurally sensitive environments.

Data operations. Teams label, enrich, link, infer, retain or use special-category data because they improve system performance

Known residue. Repeated detections are tolerated without changing sources, filters, functions or release decisions.

Outputs. The system is designed, prompted, fine-tuned or deployed to infer or disclose special-category information about individuals.

Consequence: leave the residual-data pathway. Identify an applicable Article 9(2) derogation, redesign the purpose and data flow, or stop the processing. Public accessibility is not the same as data being manifestly made public by the data subject.

Important: a separate Article 6 inquiry remains necessary. Safeguards support necessity, proportionality and accountability; they do not become the legal basis or cure deliberate Article 9 processing.

Primary-source reading map

Status note (27 July 2026): EDPB Guidelines 03/2026 are Version 1.0, adopted for public consultation on 7 July 2026; the consultation remains open until 30 October 2026. The Digital Omnibus amendment remains proposed law.

EDPB. Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0 (7 July 2026). *Paras 67-77: prohibition, four conditions, lifecycle measures and accountability/state-of-the-art review.* [Official source](#)

CNIL. The legal basis of legitimate interest: focus sheet on measures to implement in the case of data collection by web scraping (5 January 2026). See “Mandatory measures,” incidental sensitive-data collection and “Additional safeguards.” English is a courtesy translation; the French version prevails. [Official source](#)

CJEU. GC and Others v CNIL, C-136/17, judgment of 24 September 2019. *Paras 35-48, especially 45, 47-48; operative point 1.* [Official source](#)

CJEU. Meta Platforms and Others v Bundeskartellamt, C-252/21, judgment of 4 July 2023. *Paras 68-85; especially para 85 on data manifestly made public.* [Official source](#)

CJEU. Vyriausioji tarnybinės etikos komisija, C-184/20, judgment of 1 August 2022. *Paras 117-128 on indirect revelation of protected information.* [Official source](#)

EDPB. Report of the work undertaken by the ChatGPT Taskforce (23 May 2024). *Paras 13 and 18-19: derogations and filtering/deletion at scale.* [Official source](#)

European Commission / Parliament. Digital Omnibus, COM(2025) 837, procedure 2025/0360(COD). *Proposed Articles 9(2)(k) and 9(5); remains pending.* [Official source](#)

EDPB-EDPS. Joint Opinion 2/2026 on the Digital Omnibus proposal (10 February 2026). *Paras 46-52: scope, documentation and lifecycle safeguards.* [Official source](#)

This governance checklist is a practical educational resource, not legal advice. Apply it to the specific processing architecture, jurisdiction, controller roles and current version of the relevant guidance and law.