Asylum Research&Global Assistance

Web scraping and generative AI

Response to EDPB Guidelines 03/2026

European Data Protection Board

21 October 2026 | France

ARGA — Asylum Research & Global Assistance submits this position from a fundamental-rights and access-to-remedy perspective. Our central concern is the loss of context when information about identifiable people is collected, combined and reused by generative AI systems. The recommendations below concern the consultation draft and are proposals for clarification, not assertions that every suggested measure is already mandatory.

1 Make provenance operational

We support clearer accountability for the acquisition and reuse of personal data. The final guidance should connect the allocation of responsibilities in paragraphs 17–20 with a practical evidence trail. A controller should be able to link a training-data collection to its source, collection date, stated purpose, assessed legal basis, relevant restrictions and subsequent dataset versions. The level of detail should reflect risk and avoid retaining unnecessary personal data solely to document compliance.

We recommend a minimum provenance record covering the collecting entity, source categories, acquisition method, relevant filtering decisions, transfers and the training runs or model versions that used the material. For purchased datasets, a supplier warranty alone should not close the inquiry. A proportionate verification process should test the supplier's account, record unresolved gaps and identify circumstances requiring exclusion or further investigation.

The guidance should explain how responsibility follows actual decisions about purposes and means. Contractual labels must not obscure a developer's own decisions about selection, combination or reuse. An upstream breach should not automatically be attributed to every downstream actor without examining that actor's processing and applicable duties.

2 Assess lawfulness before collection

Public availability does not resolve lawfulness, necessity or fairness. Where legitimate interests are relied upon, the assessment should identify a concrete interest, examine less intrusive means and address reasonable expectations in the original publication context. A general aspiration to improve a model should not substitute for that analysis. The assessment should be revisited when the intended use or affected population changes materially.

We recommend clearer examples involving asylum-related narratives, humanitarian assistance pages, allegations of crime and information about children. These are illustrative risk scenarios, not client case studies. The same passage can acquire a substantially different impact when detached from its date, procedural status or original audience.

The guidance should distinguish special-category data under Article 9 GDPR from data concerning criminal convictions and offences under Article 10. A balancing exercise under Article 6(1)(f) does not itself satisfy those separate conditions. Controllers should identify whether such material is likely to be collected and adopt source selection, exclusion and handling measures before ingestion, with a documented response to residual collection.

3 Preserve context and enable correction

The minimisation and accuracy discussion should address outdated or contested allegations. A withdrawn notice, an acquittal or a correction can be lost when a dataset retains only an earlier

report. Where personal data are retained, provenance and update mechanisms should make the date, source and relevant status discoverable without indiscriminately expanding the data collected.

We recommend an escalation procedure for credible reports of serious inaccuracies: assess the report, identify affected dataset versions, restrict further reuse where justified, implement the applicable correction or deletion and document verification. The final guidance should distinguish action on a source corpus from action on a trained model and on generated outputs. Removing a row from a dataset does not demonstrate that the model has ceased reproducing it.

Controllers should explain the capabilities and limits of their remediation method. Technical difficulty should be evidenced, not used as a generic exemption. Tests should examine whether relevant output risks persist, while preventing the testing process from exposing the affected person's data more widely.

4 Make transparency and remedies usable

For paragraphs 24–34, we recommend concrete criteria for assessing and documenting any Article 14 exception. Scale alone should not be presented as an automatic justification. Where an exception applies, alternative information should be specific enough to help people understand the processing and locate a usable contact point.

A searchable source register can be useful, but a public register of exact URLs may itself expose vulnerable people. The guidance should permit layered disclosure: meaningful public information on source categories and periods, a protected request route, and fuller evidence available to competent authorities. It should not require fresh identifying data solely to make an anonymous or unidentifiable record attributable to a person.

Rights procedures should accept reasonably sufficient evidence and provide a reasoned outcome. A request should be traceable across the collecting entity, dataset supplier and model provider where relevant responsibilities exist. Referral between organisations must not become an endless loop. Any communication of rectification or erasure should follow the applicable GDPR requirements and protect the requester's identity.

5 Require evidence proportionate to risk

ARGA recommends a concise accountability annex covering provenance, the lawfulness assessment, source exclusions, impact assessment where required, rights handling, remediation tests and material changes. For particularly consequential processing, suitably independent review can improve confidence; the guidance should explain when such review is useful without presenting universal external certification as a legal requirement.

Reviewers should have relevant expertise, access sufficient to test claims, and safeguards against conflicts of interest. Published findings should describe scope and limitations. Neither an audit label nor a supplier assurance should create immunity from responsibility. Our priority is verifiable prevention and effective correction, rather than documentation that cannot be tested.

References

[1] EDPB Guidelines 03/2026, consultation draft, version 1.0, adopted 7 July 2026. Principal references: paragraphs 17–20, 24–34 and the sections on minimisation, accuracy and lawfulness. Official source

[2] Regulation (EU) 2016/679, in particular Articles 5, 6, 9–11, 14, 16–19, 21, 24, 25 and 35. Official source