Comments on Guidelines 03/2026: web scraping and generative AI

Zigmārs Ancveirs | Latvia | Independent cybersecurity and technology researcher | 11 October 2026

These comments concern the practical application of the draft Guidelines 03/2026 (version 1.0, 7 July 2026). The draft already addresses legal bases, transparency, minimisation and the handling of sensitive data. My suggestions focus on the evidence an organisation should be able to produce when those safeguards are questioned.

1. Purchased datasets and responsibility — paragraphs 16–20

Paragraph 19 distinguishes the collector's processing from the subsequent use of a dataset by an AI developer. A short example of the checks expected from a developer buying scraped data would be useful. Before reuse, the buyer could establish who collected the data, which source categories and periods were covered, what collection and exclusion rules applied, and how objections or corrections will be passed on. Supplier assurances cannot replace the buyer's own assessment of the lawful basis and purpose of its processing. The depth of verification should depend on the risks of the dataset.

2. Source transparency without further exposure — paragraphs 24–34

Paragraphs 30–32 already recommend public descriptions of sources and, where possible, searchable URLs and collection dates. The guidance could distinguish between information suitable for publication and records retained for supervisory examination. A public notice could identify source categories, collection periods, intended purposes and a contact route for exercising rights. A protected internal inventory could record material source decisions, filtering outcomes, removals and dataset revisions. Publishing that inventory in full may recreate a directory of personal data or expose sensitive infrastructure. Conversely, difficulty listing every URL should not excuse the absence of any auditable record of collection.

3. Website signals are not legal authorisation — paragraphs 37, 45 and 64

The draft explains that the absence of robots.txt does not constitute GDPR consent. It would be worth adding the converse point: following a robots.txt or ai.txt instruction, or passing a website access check, does not establish a lawful basis for processing personal data. These signals may be relevant to collection restrictions and reasonable expectations, but their meaning differs from the assessment required by Articles 6 and 9 GDPR. Copyright permissions and contractual access conditions also need separate consideration. A concrete example would help avoid treating a machine-readable signal as a general permission to process data.

4. Requests received before and after training — paragraphs 29, 33, 60 and 66

The discussion of objections and post-training limitations would benefit from a staged example. A request may concern data not yet collected, a retained source record, a prepared training dataset or information potentially reproduced by a trained model. Those situations cannot always be remedied in the same way. Organisations should describe what they can actually remove, what they can only suppress at output, and what remains technically uncertain. A refusal or partial remedy should be recorded and explained. The example should also preserve the Article 11 GDPR distinction: a person should not have to disclose additional identity data merely to discover whether an organisation holds their information.

5. Testing controls for incidental sensitive data — paragraphs 67–77

The treatment of GC and Others (C-136/17) is appropriately conditional; it should not be read as a general Article 9 exemption. More practical detail would help controllers demonstrate that incidental collection has been limited. Evidence might include the categories excluded, tests of collection filters, sampled findings from collected material, removal procedures and checks for model reproduction of sensitive information. False negatives and test limitations matter as much as nominal detection rates. If significant sensitive data continue to appear, the controller must reconsider the relevant processing and legal assessment, rather than rely on an

output filter as a complete remedy.

6. Demonstrating necessity and accuracy — paragraphs 35–66

A worked example of an Article 6(1)(f) assessment would make the guidance easier to apply. It could show a defined training purpose, less intrusive alternatives considered, collection limits, foreseeable effects on individuals (especially children), and the evidence used to judge whether safeguards work. This need not be a mandatory checklist for every project. The discussion of accuracy should also keep two tests distinct: whether a collected record was accurate and current when used, and whether a deployed model produces accurate statements about an identifiable person. A reliable source and an extraction timestamp do not establish either point on their own.

The common thread is accountability that can be examined in practice. Organisations should be able to explain their collection decisions and remediation limits without collecting additional personal data simply to prove compliance.

Reference: EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adopted for public consultation on 7 July 2026. Paragraph numbers refer to that draft.