# Public consultation on EDPB Guidelines 03/2026 on the processing of personal data in the context of web scraping to develop generative AI models

**From:** Vasanth Rajendran, individual expert contributor (personal capacity; views are my own)

---

From: Vasanth Rajendran, individual expert contributor (personal capacity; views are my own)

Dear EDPB,

I welcome these Guidelines and submit the following comments as an individual expert. I am a Software Development Manager at Amazon where I lead a team shipping production agentic AI systems, and I serve as a Privacy Bar Raiser with sign-off authority on privacy design for product launches. I chair the IEEE P1968 working group developing a recommended practice for the governance of autonomous AI agent systems (IEEE standards project approved September 2026), I developed the SAFE-AGENT-L governance framework presented at the AAAI-26 AI and Law Bridge workshop, and I provided expert input to NIST's TEVV-Athlon work (AI 200-2). My comments focus on making the Guidelines' balancing test operational for large-scale scraping, drawing on production privacy-review practice.

1. Give the legitimate-interest balancing test operational criteria, require it documented before collection, and wire it to the mitigations (Sections 3.2.2 and 3.2.3, paras 35-38 and 46-66).

The draft's three-part legitimate-interest test is thorough, but controllers need to know what a passing assessment looks like in practice. I recommend four additions. First, operational criteria for the balancing step: collection scale, data sensitivity, data-subject expectations, and the specific mitigations applied, each weighed explicitly. Second, timing: the balancing assessment, with the mitigations it relies on, should be documented before collection begins, not reconstructed afterward; this is the only point at which the assessment can actually shape the operation, and it evidences the accountability principle in para 15. Third, a worked example mapping mitigations to balancing factors (for instance, how pre-collection filtering and opt-out honoring reduce the weight of the "context

of processing" and "reasonable expectations" factors in paras 57-64). Finally, the Guidelines should state that a downstream re-user of an already-scraped dataset (the second scenario in paras 1-6) must conduct its own balancing assessment reflecting its own purposes and mitigations, and cannot rely on the original scraper's documentation alone.

2. Make data minimisation auditable across the whole pipeline, before and after collection (Section 2, p. 7; paras 35-38).

Pre-collection, I recommend requiring auditable data-flow mapping: a documented record of why each collected category is necessary for the stated purpose, reviewed before crawling begins. On filtering technique, the draft's mention of syntax-based filtering such as regular expressions (para 37) should be framed as the floor, not the ceiling: classifier-based filtering is the current production standard for identifying personal data categories at scale, and the Guidelines should say so. Post-collection, the draft correctly notes that untargeted scraping leaves the controller with limited knowledge of what was collected (p. 7); at web scale, pre-collection precision is impossible, so compliance must be demonstrated through post-collection classification, filtering, and deletion pipelines with documented coverage and residual risk. Together, auditable pre-collection mapping and expected post-collection filtering give controllers a coherent minimisation methodology rather than a menu of examples.

3. Give machine-readable opt-outs clear normative weight (paras 38, 62-64; Examples 4 and 5).

The draft lists robots.txt, ai.txt, authentication, and CAPTCHA as relevant signals, and Examples 4 and 5 correctly treat platform prohibitions as shaping reasonable expectations. I recommend going one step further: state explicitly that systematic disregard of a clear machine-readable opt-out weighs heavily against the controller in the balancing test, while good-faith honoring of such signals, including text-and-data-mining reservations, weighs in the controller's favor. Machine-readable signals are the only scalable expression of publisher expectations, and AI training pipelines already implement them at scale; clear normative weight creates an auditable incentive aligned with the draft's own logic.

4. Address the consequences of non-deletability for data subject rights (para 60).

The draft notes that once a model is trained, personal data cannot be easily deleted from the model. This has direct consequences for Articles 17 and 21 that should not remain implicit. If erasure from the trained model is not technically feasible, controllers need a documented position before training begins. I recommend requiring controllers to document, as part of their DPIA and balancing-test record, their technical approach to post-training rights requests: available unlearning or editing techniques, output gating and filtering, dataset versioning with a defined retraining cadence, and, where none suffice, a frank assessment of residual risk. Deletion must be treated as a lifecycle problem spanning data and model, not a dataset-only problem.

5. Require operational detection and purging of incidentally collected special-category data (Section 3.2.4, pp. 19-23).

At web scale, incidental collection of special-category data is a statistical certainty, not an edge case. I recommend the Guidelines require continuous, operational handling: post-collection classifier-based detection of special-category data, purging on a cycle matched to the scraping cadence, and documentation of both in the DPIA. This "detect, purge, document" loop should run for the lifetime of the scraping operation, not as a one-time exercise. It is the only approach proportionate to a risk the draft correctly identifies as inherent in large-scale collection.

I thank the EDPB for this important work and for the opportunity to comment.

Respectfully, Vasanth Rajendran (personal capacity)