Comments on Guidelines 03/2026 on Web Scraping in the Context of Generative AI

Submission to the European Data Protection Board | Public consultation | 29 October 2026
Submitted in a personal capacity by an individual based in Japan

Summary of recommendation

I welcome the EDPB's guidance on the risks posed by large-scale collection of personal data for generative AI. I respectfully recommend that the final Guidelines expressly distinguish (a) lawful collection and data minimisation, (b) protection of lawfully retained training datasets against compromise, and (c) prevention of personal-data memorisation, regurgitation and inference by AI models. Each requires separate controls. In particular, the Guidelines could recognise distributed data-protection methods as possible supplementary safeguards for the second category, without presenting them as substitutes for the first or third.

1. Paragraphs 34 and 35–38: security-by-design throughout the data lifecycle

Paragraph 34 refers to technical and organisational measures ensuring a high level of security; paragraphs 35–38 address minimisation before, during and after collection. I support the draft's emphasis on source exclusions, collection filters, synthetic data, deletion, anonymisation and pseudonymisation.

I suggest adding a short clarification: where personal data must lawfully remain in a dataset for a defined period, controllers should also consider post-compromise exposure minimisation. For example, they may evaluate architectures that separate intelligible data, reconstruction information and administrative control across independently protected trust domains, where this is appropriate to the risk and operational requirements.

These measures should complement—not replace—purpose limitation, minimisation, access control, encryption, logging, retention limits, deletion and effective data-subject rights.

2. Paragraphs 65–66: broaden the illustrative risk-mitigation examples

Paragraphs 65–66 discuss the role of mitigating measures in the Article 6(1)(f) legitimate-interest assessment and give a non-exhaustive list, including excluding sensitive sources, transparency, deletion, anonymisation, pseudonymisation and controls against model memorisation or regurgitation.

I propose that the illustrative list also mention, where proportionate, technical controls intended to reduce the amount of intelligible personal information exposed when a data repository, privileged account, cloud service or processor is compromised. The relevant outcome could be assessed by threat modelling, adversarial testing and evidence of exposure reduction under a single-point-compromise scenario.

This should not imply that security safeguards can legalise an otherwise unlawful scraping activity or automatically satisfy the balancing test. Measures required under other GDPR obligations must not be double-counted as additional mitigation, consistent with paragraph 65.

3. Technology-neutral recognition of supplementary protection methods

Secret sharing and all-or-nothing transform (AONT) methods are distinct technologies that may, in suitable designs, contribute to data-at-rest protection or reduce the usefulness of data obtained from an isolated compromise. They are not synonymous with anonymisation or pseudonymisation, and their effectiveness depends on architecture, reconstruction thresholds, key handling, attacker capabilities and the availability of other data. They do not prevent a fully authorised processing environment from accessing reconstructed plaintext.

As one illustrative example of a commercially available distributed data-protection approach, ZenmuTech (Japan) offers ZENMU Virtual Drive. Its mention is illustrative only, not an assertion that the product has been validated for a particular AI training pipeline, and not a request for regulatory endorsement or procurement preference. The EDPB should remain technology-neutral and focus on objectively verifiable security outcomes.

4. AI-specific risks must be addressed separately

The draft rightly raises the risks of training-data memorisation, regurgitation and model inference (including paragraph 66(i)). Securing the underlying dataset does not by itself prevent an AI model from memorising or reproducing personal information. Appropriate model-level assessment and controls—including data deduplication, training-data governance, extraction testing and output safeguards—remain necessary.

5. Suggested clarification for the final Guidelines

“Where personal data are lawfully retained for AI development, controllers should consider, in addition to access control, encryption, data minimisation and retention limits, proportionate measures to minimise the exposure of intelligible data following a security compromise. Technology-neutral examples may include separation of data and reconstruction information across independently controlled environments, where appropriate. Such measures do not themselves establish the lawfulness of collection, achieve anonymisation, or eliminate model-level risks such as memorisation and regurgitation.”

Thank you for considering this recommendation.

Reference: EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0 (7 July 2026), especially paragraphs 34–38 and 65–66.