Response to EDPB Guidelines 03/2026 on web scraping in the context of generative AI

I am responding as an individual. I welcome the EDPB's work to clarify data-protection requirements for web scraping. My central concern is that people should not have to withdraw from public life online to protect their personal information from uses they neither chose nor reasonably anticipated.

The final guidelines should preserve their existing safeguards while making the following points clearer and more practical.

1. Public accessibility should not establish permission or reasonable expectations of AI training

Relevant provisions: paragraphs 45 and 63-64; Example 4.

I welcome the clarification that making personal data publicly accessible does not constitute consent to scraping, and that the absence of a robots.txt restriction is not consent.

However, Example 4 risks giving excessive weight to a platform's warning that scraping may occur. Please clarify that such a warning, combined with public accessibility and an absence of technical restrictions, cannot by itself establish reasonable expectations of generative-AI training.

Knowing that an activity might happen is different from reasonably expecting one's information to be used for a particular purpose. People may publish information to seek support, participate in public debate or communicate with a community. They may have little control over the platform's technical settings.

Please distinguish more clearly between data published by the person concerned and information published about them by someone else. Neither a platform nor another user should be treated as expressing that person's consent.

2. The treatment of incidental sensitive-data collection needs tightly defined limits

Relevant provisions: paragraphs 67-77.

The guidelines acknowledge special protection for sensitive personal data and state that the search-engine analogy is not a general exemption. I support that limitation.

Please explain more precisely which similarities with search-engine processing would justify applying the analogy to AI training, and which differences would prevent it. In particular, the assessment should address whether sensitive information can be retained, reproduced or inferred after the original training material has been removed.

Please require controllers to document why collection is genuinely incidental and residual, what prevention measures were tested, and how failures are detected and corrected. Repeated collection from a known high-risk source should trigger reassessment rather than continuing to be characterised as unavoidable.

The final text should also specify when ineffective safeguards require suspension of the affected processing. Output filtering should not be presented as equivalent to deletion or as sufficient merely because sensitive information is less visible to users.

Please include practical examples covering children's information and people exposed to discrimination, abuse or other heightened risks, including where their information appears on general-purpose websites.

3. People need usable routes to exercise their rights

Relevant provisions: paragraphs 29-33, 60 and 66(f).

Please set out a practical minimum standard for requesting access, correction, objection and deletion in this context.

A person should not be expected to identify an unknown scraper, reconstruct a training dataset or prove that a model contains their information before receiving meaningful assistance. Controllers should provide a clearly identifiable contact route, explain what information is reasonably needed to locate relevant records, and avoid demanding excessive additional personal information.

Requests should produce an understandable account of what was searched, what action was taken and any limitations. Please distinguish action affecting source datasets, retained copies, future training and already trained models.

The guidelines should also clarify how discretionary pre-collection opt-outs relate to existing rights. An unused opt-out mechanism should not be treated as consent or as resolving the controller’s separate obligation to establish lawfulness.

4. Large-scale collection should not automatically excuse individual notification

Relevant provisions: paragraphs 25-34; Example 2.

Please strengthen the explanation of how controllers must justify reliance on disproportionate effort. The assessment should consider the consequences of people remaining unaware, including particularly serious effects on smaller groups within a large dataset.

Please explain how a dataset-wide assessment can avoid overlooking those groups. Where individual notification is legitimately unavailable, information should be discoverable by affected people in practice, rather than merely present somewhere on an unfamiliar company’s website.

5. Necessity and safeguards should be demonstrated, not assumed

Relevant provisions: paragraphs 35-38, 49-52 and 65-66; Example 7.

Please ask controllers to document comparisons with less intrusive approaches, including narrower sources, smaller datasets and approaches that avoid personal data. The explanation should connect the data collected to a sufficiently specific purpose, rather than relying on a general ambition to improve AI.

Please qualify Example 7 so it cannot be read as a checklist providing automatic approval. Copyright permissions, public accessibility and technical safeguards should not substitute for the separate data-protection assessment.

I support final guidelines that place responsibility on organisations to demonstrate lawful and proportionate processing, rather than leaving individuals to discover and contest uses of their information after the consequences have become difficult to reverse.