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

Reply of the OpenMined Foundation to the public consultation

October 2026

1. Who we are

OpenMined is a non-profit foundation and an open-source community of more than 17,000 engineers, researchers, and students. Since 2017 we have built free, open-source software around a simple idea: organisations should be able to answer questions from data they cannot see. In the systems we build, data stays with its owner, who decides what it may be used for, on what terms, and for how long. Much of our current work applies this idea to artificial intelligence, so that the people and organisations who hold data can take part in AI development without handing their data over.

We submit these comments from that practical experience. Our interest is not commercial: all of our software is free and open source.

2. Summary of our comments

The guidelines answer a hard question: when may an AI developer collect personal data from the public internet without asking the people concerned? We support the structure of the answer: legitimate interest as the realistic legal basis, a strict necessity test, a balancing test that takes people's expectations seriously, and a list of mitigating measures.

Our comments make one central point and apply it four ways. The point is that an alternative to scraping exists and is growing: permissioned access, in which data stays with the person or organisation that holds it and is used for AI development only on terms the holder sets and can withdraw. The draft, as it stands, never mentions this alternative. Naming it would sharpen the necessity test, give the balancing test a positive benchmark, and point developers toward the practice that protects people best.

- Comment 1 asks that the necessity test name permissioned access among the less intrusive means (paragraphs 51–52).
- Comment 2 asks for recognition of AI designs that keep personal data out of the model, because they restore the control that the guidelines say people lose (paragraphs 54, 60, 62–64, and 66).
- Comment 3 proposes two additions to the list of mitigating measures (paragraph 66).
- Comment 4 supports the recommendation on synthetic data, with one caveat (paragraphs 37 and 52).

We close with a note on what naming these alternatives would mean for smaller European developers.

3. Comment 1: the necessity test should name the alternative (paragraphs 51–52)

Under the guidelines, a developer relying on legitimate interest must show that there is no "equally effective and less intrusive way" of pursuing the purpose. Paragraph 52 offers three ways to reduce intrusion: narrower collection criteria, pseudonymised data, and synthetic data.

We ask the EDPB to add a fourth: obtaining the data through channels where the holder has said yes. Such channels exist and are multiplying. Publishers license their archives for AI training. Data intermediaries under the Data Governance Act organise access on behalf of groups of holders. Open-source software, including ours, lets any organisation open its data to AI development on its own terms, with each use permitted or refused at the source. Data obtained this way answers the very question the balancing test struggles with: the developer does not have to speculate about what the data subjects would expect, because the source has stated its terms.

The test should also look forward. Paragraph 77 already obliges controllers to follow the state of the art in protective measures and to adapt as it advances. The same logic belongs in the necessity test. As permissioned channels mature for a category of data, scraping that same category without permission becomes correspondingly harder to justify as necessary.

Our suggestion. In paragraph 52, add sourcing through permissioned channels, where such channels are reasonably available for the data in question, to the list of less intrusive means, and state that controllers should reassess their availability over time.

> **EXAMPLE IN PRACTICE**
>
> A developer fine-tuning a model on news text has two routes. It can scrape news sites, which also sweeps up reader comments carrying the names and opinions of people who have no idea their words are becoming training data. Or it can connect to the access points that a growing number of publishers offer, taking the articles on agreed terms, with reader comments excluded at the source. The second route yields cleaner data and an arrangement that can be corrected or ended. A scraped copy, once taken, can be neither.

4. Comment 2: give people back an objection that works (paragraphs 54, 60, 62–64, and 66)

The guidelines identify, correctly in our view, the deepest harms of scraping for AI: people lose control over data they published for another purpose (paragraph 54), and once a model has been trained, "personal data cannot be easily deleted from a model" (paragraph 60). After that point, the right to object becomes a request that the technology cannot honour.

That dead end follows from one design choice: blending all collected data into the model's parameters. Designs that avoid it exist. In retrieval-based and federated systems, the source material stays where it is, and the model consults it at the moment of answering, under rules the source controls. When a person deletes their material, or a platform withdraws from participation, future answers simply stop drawing on it. The objection becomes a switch that works.

Paragraph 66(i) already lists measures that limit memorisation and regurgitation after the fact. We ask the EDPB to name the stronger option in the same list: architectures in which personal data is never absorbed into the model at all. In the balancing test, preventing a harm by design should weigh more heavily than filtering its symptoms afterwards.

Our suggestion. Add to the list in paragraph 66: choosing model architectures that keep personal data outside the model and under the source's control, so that withdrawing access takes effect on all of the model's future outputs.

> **EXAMPLE IN PRACTICE**
>
> Someone discovers that forum posts they wrote ten years ago can be coaxed out of a popular chatbot. If the posts went into the model's training, deleting them from the forum changes nothing, and the developer can offer little beyond output filters. In a federated design, the posts would have remained at the forum. The chatbot would have consulted them only while permission stood, and the person's deletion, or the forum's withdrawal, would have taken real effect on every answer given from that day on.

5. Comment 3: two additions to the mitigating measures (paragraph 66)

First, where permissioned channels exist for part of the data a developer seeks, using them for that part belongs on the list of mitigating measures, for the reasons set out in comment 1. A developer who takes licensed news text through the publisher's access point, and scrapes only what no channel offers, intrudes less than one who scrapes everything.

Second, the guidelines treat robots.txt, ai.txt, and CAPTCHA as signals that scraping is unwelcome (paragraphs 37 and 64). We suggest recognising the positive counterpart as well: machine-readable terms through which a website states what it does allow, for example that articles may be used for AI training with attribution while photographs may not. Honouring such terms in both directions, the refusals and the permissions alike, gives websites a richer vocabulary than yes or no, and gives the people behind them a fuller expression of their choices.

Our suggestion. Add both measures to the list in paragraph 66.

6. Comment 4: synthetic data, with one caveat (paragraphs 37 and 52)

We support the recommendation to consider synthetic data before and during collection. One caveat belongs in the text: synthetic data is only as protective as the way it was generated. A generator trained carelessly on personal data can reproduce it, a point the EDPB's own draft Guidelines 02/2026 on Anonymisation make in their paragraph 82. The recommendation will achieve its aim if it reads "synthetic data generated with appropriate privacy guarantees", ideally with a cross-reference to the anonymisation guidelines on what appropriate means.

Our suggestion. Qualify the references to synthetic data in paragraphs 37 and 52 accordingly, and cross-reference the two sets of guidelines.

7. A closing note on who can comply

Strict duties with no named lawful path tend to favour the organisations best resourced to absorb legal risk. If the final guidelines name the alternatives, permissioned channels, source-controlled architectures, and synthetic data with guarantees, they will describe a path that a university laboratory, a start-up, or a public-interest project can actually walk, using open-source tools that already exist. We believe that serves both the people the GDPR protects and the health of European AI development.

We thank the EDPB for the opportunity to comment, and we offer our community's technical expertise in support of the final text. The OpenMined Foundation can be reached through the contact details provided with this submission.