Contribution to the EDPB public consultation

Guidelines 03/2026 on web scraping in the context of generative AI
Version 1.0, adopted on 7 July 2026 for public consultation

Subject: Distinguishing collection prevention from measures applied after collection - comments on paragraphs 36-38.
Contributor: Lajos NAGY, independent founder of Zero Data Protocol (ZDP), Belgium
Website: https://zdp.ai
Date: 26 October 2026

Contributor and purpose

Zero Data Protocol (ZDP) is an independent architectural framework developed by Lajos NAGY in Belgium. It explores reducing the processing of personal data through four principles: Zero Collection, Zero Retention, Zero Exploitation and Human Freedom Choice. This contribution proposes architectural clarifications. It claims no status as a recognised standard, certification or guarantee of GDPR compliance.

Architectural issue

The distinction between measures before, during and after collection could be accompanied by guidance on documenting their actual effects. Four situations should be distinguished: (1) not acquiring unnecessary personal data; (2) acquiring data and subsequently producing an anonymised version; (3) acquiring data and subsequently deleting it; and (4) retaining data while restricting its uses, including use for training.

These situations can serve different needs and be combined. They do not, however, have the same effects on initial exposure, intermediate copies or subsequent uses. Our architectural recommendation is to examine reasonable opportunities to avoid particular data before acquisition, then apply appropriate safeguards to data that must actually be processed. This assumes neither that all collection is avoidable nor that post-collection measures are unnecessary.

Hypothetical illustration - not tested

An organisation intends to compile technical documentation to fine-tune a generative model. One architecture downloads complete selected pages, including comments, signatures and authors' contact details, then removes those elements from the training corpus. Another excludes discussion areas whose URLs can be recognised in advance and which are unnecessary for the purpose, then limits extraction from selected pages to relevant sections.

The second approach could avoid some acquisitions. It would not demonstrate an overall absence of collection: content received for analysis, URLs, logs, caches and processing by service providers must be examined separately. A filter that receives data before deleting it has not prevented its initial processing.

Proposal 1 - Document alternatives before collection (paragraph 36)

We suggest clarifying that the prior assessment can identify the categories of data necessary for the defined purpose, those that can reasonably be excluded, the alternatives examined and their limitations, and the reasons why certain personal data remain necessary. Proportionate documentation would make the architectural choice explainable without prescribing a single technical method.

Proposal 2 - Specify where filters operate (paragraph 37)

We suggest distinguishing exclusions that prevent a request to a source, restrictions applied upon receipt or extraction, and removal after recording. For each measure, it would be useful to indicate which data are actually received, by which entity and at what stage. This would prevent exclusion from the final corpus from being described as an absence of collection throughout the processing chain.

Proposal 3 - Include intermediate processing and copies (paragraph 38)

We suggest supplementing the documentation of measures with a proportionate examination of processing memory, caches, logs, backups and copies held by processors. For each relevant location, this examination could specify the categories of data, their utility, access arrangements and the conditions for retention or deletion.

Claims about safeguards should also specify their scope. A no-training commitment concerns a use; data residency concerns a location; and a policy described as Zero Data Retention must be assessed against its scope and exceptions. None of these elements alone demonstrates an absence of collection or successful anonymisation.

Scope and limitations

These proposals are architectural recommendations, not additional legal obligations or an exhaustive interpretation of the GDPR. Their purpose is to facilitate the description and verification of the measures selected. Prevention must remain suited to the purpose, technical capabilities and applicable obligations. Necessary data may still require security measures, use restrictions, anonymisation or deletion.

The principle "No Data to Protect" expresses a direction here: where an entity does not actually acquire unnecessary personal data, it avoids risks associated with holding those data. This statement is limited to the demonstrated scope and does not remove the need to examine other processing in the chain.

Reference

EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0, adopted 7 July 2026, paragraphs 36-38, numbered page 12.
https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf