Source traceability and correction records

Comments on EDPB Guidelines 03/2026 - Version 1.0

Artashes Nazaryan | EMID, independent project in development | Argentina

3 October 2026

Scope of this contribution

I submit these proposals from a documentary design perspective. EMID explores links between statements, sources and revisable chronologies, using fictional demonstrations and human review. It is an independent project in development. These comments concern documentation for web-scraped datasets used to train or fine-tune generative AI; they do not claim empirical validation, institutional endorsement or legal compliance.

1. Separate the dates that a source record can establish

Suggested insertion: paragraph 31; cross-reference paragraph 42.

Suggested wording: Where appropriate, source records should distinguish the collection date from the original publication or update date, when available, and from the date of the last documented verification. Missing dates and the basis for each recorded date should be made explicit. Collection time alone should not be presented as evidence that the source content is current.

Practical purpose: an old webpage collected today and an updated webpage collected today can otherwise appear equally recent. A compact record could use separate fields for publication/update, collection and verification, with an explicit unknown value where needed.

2. Preserve a proportionate path from source to training dataset

Suggested clarification: paragraphs 37-38; cross-reference paragraph 42.

Suggested wording: The inventory could link a source reference to a dataset item or batch, record material filtering or transformation steps, and identify the dataset version and review status. Such records should be proportionate, access-controlled and retention-limited. They should support verification without requiring indefinite storage of complete source pages or unnecessary personal data.

Practical purpose: a reviewer should be able to determine whether a discrepancy arose in the source, extraction, cleaning or later transformation. A checksum can support file-integrity checks, but does not establish factual accuracy, a lawful basis or authority of the source.

Operational clarifications

3. Make correction handling explicit before training

Suggested clarification: paragraphs 39-42; cross-reference paragraphs 33 and 66(f).

Suggested wording: A documented correction workflow could connect a reported discrepancy to the affected item or batch, the review outcome, the relevant dataset versions and any exclusion or correction before training. Material unresolved conflicts should be recorded rather than silently reconciled. The status and scope of a remedy should distinguish changes to source records, training datasets and model outputs.

Practical purpose: recording a request, reviewing it and implementing a correction are separate events. A correction log should avoid implying that a change to a dataset has already removed information from a trained model. It should also avoid retaining unnecessary identifying details about the person making a request.

4. Distinguish fictional examples from synthetic data derived from real data

Suggested clarification: paragraphs 37-38.

Suggested wording: When discussing synthetic data, specify whether the data are wholly fictional examples or are generated using information about real persons. Document the generation method, intended use and relevant privacy evaluation. The description synthetic should not, by itself, be treated as a conclusion that the data are anonymous or that privacy risks have been eliminated.

Practical purpose: wholly fictional examples can illustrate a documentation workflow without exposing a real case. Where synthetic data are derived from real records, the assessment should consider the actual method and possible links to individuals. These are distinct situations and should be described separately.

Reference

Reference: European Data Protection Board, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0, adopted 7 July 2026 for public consultation. Paragraph numbers above refer to that version. All suggested wording is proposed by the contributor.

Official consultation:
https://www.edpb.europa.eu/public-consultations/guidelines-032026-on-web-scraping-in-the-context-of-generative-ai_en