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

Submitted by: Meem Arafat Manab

Affiliation: Marie Skłodowska-Curie Doctoral Fellow, Ontology Engineering Group, Universidad Politécnica de Madrid (HARNESS Doctoral Network, Horizon Europe); Research Associate (pro bono), School of Religion, Theology, and Peace Studies, Trinity College Dublin

Consultation: Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0, adopted 07 July 2026

Summary

These comments address three gaps, each arising from a boundary the Guidelines draw around themselves.

First, the Guidelines are titled for generative AI but scoped to training and fine-tuning. They do not address web scraping performed at inference time, that is, live retrieval through tool-use protocols such as the Model Context Protocol, function calling, or agentic browsing, which is now a standard architecture for grounding generative AI outputs and which the analytical framework in Section 3.2 does not fit.

Second, the exclusion of data brokers at paragraph 4, read together with the allocation of responsibility at paragraph 19, leaves a configuration in which bulk scraping can be conducted by an entity outside these Guidelines' scope and reused by an entity that did not perform the collection, with the original collection never assessed under this framework by anyone.

Third, conditions (a) to (c) of the *GC and Others* test at paragraph 73 are described inconsistently across paragraphs 73, 74 and 75, and condition (c) in particular depends on a state of technical capability that is changing. As detection tooling improves, the condition becomes harder to satisfy, meaning controllers relying on it today may find that reliance lapses without any change in the law or in their own conduct.

Comment 1: The Guidelines do not address inference-time web access by generative AI systems

Location: Paragraphs 1 to 6 (scope); paragraph 7 (definition); paragraphs 12, 16, 17; Section 3.2 generally; paragraphs 40, 54, and 74(iv)

Issue:

The Guidelines define their subject matter as scraping performed to construct training corpora. Paragraph 1 frames the purpose as clarifying the implications of web scraping for generative AI "development". Paragraph 2 situates scraping "in the context of AI training", used to collect large datasets "which may be employed not only to train new generative AI models but also to fine-tune existing models for specific use cases". Paragraph 3 sets out two typical situations, both concerned with dataset construction: an organisation scrapes or contracts scraping in the context of the development of generative AI, or obtains and reuses a dataset already scraped by a data broker. Paragraph 5 excludes processing of an organisation's own personal data, as against data acquired from external sources.

None of these describes a generative AI system that, in the course of responding to a user, issues a tool call to retrieve a live web page and incorporates its contents into the response. Whether that activity falls within the definition at paragraph 7 is unclear on the face of the text. Paragraph 7 defines web scraping as the use of automated tools "for extracting and storing information from publicly available web services". A retrieval that is passed into a context window and discarded may not involve storing, although retrieved content is often cached, logged, or retained for evaluation, and paragraph 12 lists extraction, cleaning, structuring, and storing as alternative processing operations rather than cumulative ones. The activity is generative AI; it is automated extraction from publicly available web services, and it is neither training nor fine-tuning.

Retrieval-augmented generation and tool-using deployments are now standard architecture in generative AI products. The Guidelines' silence on them is therefore consequential, and it is not clear from the text whether it is deliberate.

It is not true that the Guidelines are scoped uniquely to the training phase. Paragraph 40 reaches the model once it is placed on the market. Paragraph 74(iv) requires monitoring of output after development. Paragraph 54 contemplates a model connected to a web search tool, treating that connection as a factor aggravating the “chilling effect” on data subjects who intend to express themselves anonymously. A live web-connected model therefore appears in these Guidelines as a risk factor within another controller's balancing test, but never as an entity performing extraction in its own right. The boundary is drawn at collection, and it is drawn inconsistently.

The framework established in Section 3.2 (“Guidance on specific provisions in the GDPR”) was built for collection decided in advance, and does not transfer cleanly to retrieval triggered by a user's prompt, as we explore below.

- **Controller allocation (paras 16 to 20).** The roles contemplated, namely AI developer and contracted scraper, presuppose a discrete corpus-building exercise with a documented instruction set (para 18). Where a system selects a URL at runtime in response to a user's prompt, there is no such instruction set, and paragraph 17's observation that the organisation performing the scraping is not necessarily the controller becomes difficult to apply. Paragraph 16 directs a case-by-case analysis of the factual circumstances. Still, the factual circumstances here differ in kind from those the Guidelines describe: the purpose is set by the user at the moment of the request, not by the developer in advance.

- **Transparency (paras 24 to 34).** The Article 14(5)(b) disproportionate effort analysis is built for the fact pattern in Example 2: data of thousands or millions of individuals, drawn from sources extending back twenty years, with no direct link between scraper and data subject. A single live fetch presents close to the inverse, with an identified source, a narrow and often traceable set of data subjects, and a much lower cost of notice or exclusion. Paragraph 28's instruction that the balancing be conducted "with regard to the dataset as a whole and not on every single piece of personal data" has no obvious application where there is no dataset. Paragraph 31's good practice of recording domain names, URLs and dates of collection would in this context produce a log of user activity, which raises its own questions.

- **Necessity and balancing (paras 51 to 66).** Paragraph 52 advises narrowing collection criteria to satisfy necessity. For a query-triggered retrieval, there are no standing collection criteria to narrow, and necessity would have to be assessed per request against a purpose defined by the user's prompt. The reasonable expectations factors at paragraph 64, including the nature of the source website, restrictions such as robots.txt or ai.txt, and the characteristics of the data subject, remain relevant but were calibrated for a controller deciding in advance what to crawl, not for a system resolving a URL mid-conversation.

- **Special categories (paras 67 to 77).** See Comment 3 below.

Proposed amendment:

State in Section 1 whether inference-time, tool-mediated retrieval falls within scope, and clarify whether the storage element of the paragraph 7 definition is intended to

exclude it. If it falls within scope, add a worked example addressing it, since it diverges from Examples 1 to 7 in the respects set out above. If it does not, say so, and indicate whether separate guidance is contemplated. Given paragraph 54, it would also help to state whether a controller deploying a web-connected model is understood to be performing scraping, or only to be aggravating the risks of scraping performed by others.

Comment 2: The data broker exclusion and the re-use rule together leave collection unassessed

Location: Paragraphs 3, 4 and 5; paragraphs 14, 19, 32, 43 and 60

Issue:

Paragraph 4 excludes from these Guidelines "the role of data brokers who scrape data and just hold data sets available for other organisations, and whose intention is not to train AI models themselves".

Paragraph 19 provides that where an AI developer trains on a dataset already scraped by another entity, the two are in principle responsible for their own separate processing, and that "the scraper is not, in principle, responsible for the re-use of the data".

Read together, these produce a configuration in which:

- the entity that performed the collection is outside the scope of these Guidelines (paragraph 4) and not responsible for downstream re-use (paragraph 19); and
- the entity that trains on the data was not the collector, and is identifying a lawful basis under paragraph 43 for a re-use whose underlying collection it did not design and may not be able to reconstruct.

The Guidelines are not exactly silent on the acquiring controller's position. Paragraph 32 recommends providing the means to contact the controller from whom the data was obtained, linking to that controller's website, and explaining the conditions under which the personal data was collected. But a recommendation to signpost the original collector does not substitute for that collector's collection having been assessed against the principles in Section 3.2. On the Guidelines' own terms, no one is required to have done that.

This matters more given paragraph 14's observation that it may be challenging for both the scraping organisation and the organisation obtaining the scraped data "to determine

exactly what personal data, and related to which data subjects, are included in the data set", and paragraph 60's observation that once the model is trained, personal data cannot easily be deleted from it. Where the accountability gap sits at collection, it is not recoverable later in the chain.

Paragraph 5 closes with the statement that the exclusion of an organisation's own personal data "does not mean that the excluded situations are not subject to data protection legislations". If "the excluded situations" is read to reach paragraph 4 as well, part of the clarification sought below is already present. But the sentence sits in a paragraph about a different exclusion, and paragraph 4 carries no equivalent saving clause, so the position is not clear on the face of the text.

The exclusion is also unstable in application. The distinguishing feature at paragraph 4 is the broker's "intention" not to train models itself. Intention is self-declared and revisable, and an entity that scrapes at scale, sells datasets, and later develops models does not fall neatly on one side of the line.

Proposed amendment:

1. Confirm that the saving clause at the end of paragraph 5 applies equally to the exclusion at paragraph 4, or add an equivalent statement to paragraph 4, and indicate which obligations continue to apply to a scraping data broker.
2. Where an AI developer acquires a dataset it did not collect, provide that its assessment under paragraph 43 should extend to whether the original collection satisfied the principles in Section 3.2, and that an inability to establish this is a relevant factor in the balancing test rather than a neutral one.
3. Consider whether the intention to train is the right dividing line at paragraph 4, given that it is self-declared and revisable.

Comment 3: The conditions of the *GC and Others* test are described inconsistently, and condition (c) rests on a capability state that is changing

Location: Paragraphs 68, 72 to 77, in particular paragraph 73(c), paragraph 75 and paragraph 77

Issue:

The tests at paragraphs 73 and 75 are not the same test. Paragraph 73 directs the controller to assess whether the four conditions "are met". Paragraph 74 refers to

conditions (a) to (c) in paragraph 73 being "relevant for the specific processing activity". Paragraph 75 requires the controller to demonstrate that they "are relevant for the specific processing activity". Whether a condition is met and whether it is relevant are different questions, and the difference matters for what a controller must show under the accountability principle. The point below assumes the paragraph 73 formulation, but the Board may wish to reconcile the three paragraphs in any event.

Condition (c) in paragraph 73 rests on a changing state. It provides that the controller should assess whether "it is difficult/impossible to assess whether and to what extent the processing includes special categories of personal data", and therefore to prevent the collection of such data. It is drafted as a fixed condition, but it is in part a statement about the state of the art in classification and detection. That state is not fixed at the moment. As tooling for identifying special category content in scraped text and images improves, including through the same model capabilities these Guidelines are otherwise concerned with, the difficulty of assessing whether a corpus contains such data declines.

A distinction should be drawn here, because paragraph 68 grounds condition (c) partly in sequence rather than capability: it observes that it is not always possible to assess whether scraping has collected special category data "until after the data has been scraped, thus when the processing activity has already been carried out". That sequencing problem does not dissolve as detection improves. Better classifiers will still operate on data already collected.

What does change is the second half of condition (c), namely the ability to "therefore prevent the collection". Paragraph 74(i) requires the controller, before collection, to define precise criteria and apply filters to prevent the collection of special categories of personal data. The efficacy of those filters is a direct function of detection capability. As pre-collection filtering improves, the ground for saying that collection could not have been prevented narrows, even though the post-hoc character of full assessment remains.

The Guidelines are otherwise alert to the general point. Paragraph 77 requires that measures be in line with state-of-the-art techniques, notes that better privacy enhancing techniques may become available, and directs controllers to "constantly monitor technological developments to ensure that the state-of-the-art measures remain appropriate". Paragraph 76 requires regular verification of effectiveness.

But paragraph 77 addresses the measures a controller must implement, not the precondition at 73(c) that made the framework available in the first place. A controller may therefore be updating its measures under paragraph 77 while the condition that entitles it to rely on the framework at all has ceased to hold. Given paragraph 75's

demonstration requirement, this is a live accountability question rather than a theoretical one.

The Guidelines state at paragraph 72 that the ruling is not a general exemption and applies only to processing activities similar to those in the judgment and under similar circumstances. That limitation would be better served by making the temporal instability of condition (c) explicit.

Proposed amendment:

1. Reconcile paragraphs 73, 74 and 75 on whether conditions (a) to (c) in paragraph 73 must be met or must be shown to be relevant, and state what the demonstration under paragraph 75 requires.
2. Add to paragraph 77, or to paragraph 75, that condition (c) in paragraph 73 must be re-established over time, and that a controller's ability to rely on the *GC and Others* framework may lapse as detection and pre-collection filtering capability improves, independently of whether its mitigating measures remain adequate.
3. In doing so, distinguish the two underlying components of condition (c) in paragraph 73: the sequencing problem identified at paragraph 68, which improved detection does not resolve, and the preventability of collection under paragraph 74(i), which it does affect.

Closing note

The three gaps above share a common shape. Each arises where the Guidelines draw a boundary, whether around training as against inference, around AI developers as against data brokers, or around what is currently difficult to detect, and the activity on the far side of the boundary is neither addressed here nor evidently addressed elsewhere. The first two could be resolved by clarifying scope. The third requires recognising that one of the conditions on which Section 3.2.4 depends is not a constant.