Contribution to the public consultation

Guidelines 03/2026 on web scraping in the context of generative AI (Version 1.0, adopted on 7 July 2026)

Submitted by	Thiago Francis Alves da Silva
Capacity	Independent software architect – privacy engineering, regulatory compliance infrastructure and tamper-evident audit trails
Location	Alpes-Maritimes, France
Date	28 October 2026
Previous contributions	Guidelines 02/2026 on Anonymisation (submitted on 4 August 2026); Guidelines 04/2026 on administrative fines (submitted on 28 October 2026)
Publication	I consent to the publication of this contribution under my name.

Summary

The lawfulness of web scraping under the Guidelines depends on facts that exist only at the moment of collection and that are perishable: which restrictions a website displayed when it was accessed, which filters the crawler applied, which sources were in the list at that time. The Guidelines rightly require controllers to respect those restrictions, to implement mitigating measures and to be able to demonstrate that they did so. They do not say how such facts are to be recorded. This contribution proposes (i) a new paragraph recommending a **record of the collection context** whose date and integrity can be independently verified, (ii) consequential clarifications to paragraphs 31, 64, 66(e) and 75, (iii) a worked example in the format of the Guidelines, and (iv) a cross-reference to Article 53 of the AI Act, under which the same facts must already be established by providers of general-purpose AI models. The proposal creates no new obligation: it gives a concrete and proportionate form to the accountability principle that the Guidelines already invoke.

1. Scope of this contribution

This contribution concerns Section 3.2.2 (transparency, paragraphs 30 to 32; data minimisation, paragraph 37; accuracy, paragraph 42), Section 3.2.3 (reasonable expectations and mitigating measures, paragraphs 64 to 66), Section 3.2.4 (paragraphs 74 to 76) and, as regards the reuse of scraped datasets, paragraph 19. It is written from the perspective of a practitioner who designs the technical measures that the Guidelines ask AI developers to implement, and who has to make them demonstrable afterwards.

2. Observations on the draft

2.1 The Guidelines rest on facts that exist only at the time of collection

Across the Guidelines, the outcome of the assessment depends on the state of the source at the moment the crawler accessed it and on what the crawler did at that moment:

- **paragraph 37**: the controller should “exclude from the collection websites which clearly oppose the scraping of their content”, identified through robots.txt or ai.txt files, CAPTCHA or authentication;
- **paragraph 45**: “the absence or non-applicability of a robots.txt file on a web site does not amount to consent” – the presence or absence of the file is therefore a fact to be established, not a presumption;
- **paragraph 64**, third indent, and Examples 4 and 5: reasonable expectations turn on the restrictions a website implements; where such measures are in place, “it is less likely that they can expect the processing of their personal data by a scraping entity”;

- **paragraph 65**: where “sufficient mitigating measures are not taken, the processing cannot be based on Article 6(1)(f) GDPR” – the measures are therefore constitutive of the legal basis, not ancillary to it;
- **paragraph 66(c)**: exclusion of data behind authentication walls or served differently to crawlers than to browsers – again a fact observable only at access time;
- **paragraph 74**: filters applied “before the collection” and deletion of special categories of data “immediately after the collection or as soon as it is identified” – events in time;
- **paragraph 42**: “Timestamp the data: record when the data was scraped” – the one place where the Guidelines ask for a dated record, but for the purpose of accuracy only.

These facts are perishable. Paragraph 10 itself notes that “Websites change their structure frequently”. A robots.txt file can be added, removed or amended at any time; a platform can introduce an opt-out for AI training months after content was collected (the situation in Example 5 can arise after the situation in Example 4); an authentication wall can be added; a page can be deleted. Third-party web archives do not record what a particular crawler saw or did. Once the moment has passed, the only evidence of the state of the source and of the crawler’s conduct is the one the controller itself made at the time.

2.2 The Guidelines do not say how these facts are to be recorded or demonstrated

The Guidelines acknowledge the difficulty. Paragraph 14 states that “It is thus not obvious how to meet the principle of accountability laid down in Article 5(2) GDPR”, while recalling that compliance with Article 5(2) GDPR “must be met by the controller”. Paragraph 75 requires that “The controller should be able to demonstrate, in line with the accountability principle” that the conditions for incidental processing of special categories of data are met, and paragraph 76 calls for “regular verification and assessment” of the measures. Yet, across the 77 numbered paragraphs, the words “documentation”, “evidence”, “audit” and “provenance” do not appear; “demonstrate” appears once (paragraph 75); “documented” appears once, about the processor’s instructions (paragraph 18); “record” appears once, as a verb, for the date of scraping (paragraph 42); and “timestamp” appears only for the accuracy of the data collected (paragraph 42) and as an example of metadata that may be served to crawlers (paragraph 66(c)). The Guidelines thus require controllers to respect crawl-time facts and to demonstrate that they did, without indicating how a controller can preserve those facts in a form that a supervisory authority, a data subject or a downstream reuser can rely upon.

This has four practical consequences.

(a) The controller cannot demonstrate. A developer that genuinely honoured robots.txt and ai.txt files, excluded authenticated content and applied the filters described in paragraphs 37, 38 and 74 has, under the current text, no recommended way of proving it once the source has changed. Its position is no stronger than that of a developer that did none of these things. In Opinion 28/2024, which the Guidelines cite, the EDPB already indicated that supervisory authorities should take into account the documentation provided by controllers when assessing AI models; the present Guidelines do not say what that documentation should contain for the collection phase.

(b) The supervisory authority cannot verify. Disputes about whether a restriction was in force at the time of collection, or whether a filter was applied, become assertion against assertion. The authority is left to weigh declarations on precisely the facts that decide whether Article 6(1)(f) GDPR is available (paragraph 65).

(c) Data subjects cannot exercise rights effectively. Paragraph 60 recognises that “it might be difficult for data subjects to exercise their rights under the GDPR” and that “once the model is trained, personal data cannot be easily deleted from a model”. The prior right to object and the opt-out list in paragraph 66(f) presuppose that a data subject can learn whether, when and from which source their data was collected. Paragraph 66(e) recommends “an updated list of scraped websites”; but an updated list, by definition, overwrites the previous one. The information a data subject needs about a

past collection disappears at the next update, unless the list is versioned.

(d) Datasets travel without provenance. Paragraph 19 states that "The scraper and AI developer are, in principle, responsible for their own, separate, processing activities", and paragraph 32 recommends providing the means to contact the controller from whom a dataset was obtained. A downstream developer must therefore carry out its own balancing test under paragraph 64. It cannot do so meaningfully if the dataset arrives without the collection context: which sources, on which dates, under which restrictions, with which exclusions. The Guidelines address the reuse scenario but not the information that must accompany a dataset for the reuser to comply.

2.3 The same facts must already be established under other Union law

The facts at issue are not specific to the GDPR. Under Article 4(3) of Directive (EU) 2019/790, which Example 7 refers to, rightholders may reserve text and data mining rights by machine-readable means, in practice largely through robots.txt and comparable signals. Under Article 53(1)(c) of Regulation (EU) 2024/1689 (AI Act), providers of general-purpose AI models must put in place a policy to identify and comply with such reservations, including through state-of-the-art technologies, and under Article 53(1)(d) they must make publicly available a sufficiently detailed summary of the content used for training, according to the template provided by the AI Office. The General-Purpose AI Code of Practice commits signatories to read and follow instructions expressed in accordance with the Robots Exclusion Protocol (Measure 1.3(1)(a)) and to make publicly available information about the web crawlers they use and their robots.txt features (Measure 1.3(4)) – the same "crawler's characteristics" that paragraph 30 of the Guidelines requires controllers to publish. The Guidelines do not refer to the AI Act.

A provider subject to both regimes will have to establish, for the same crawl, which sources were accessed, when, which signals were observed and which decision was taken. If each regime is documented separately, the burden doubles and the records may diverge. A single record of the collection context, kept in a verifiable form, serves the GDPR assessment under paragraphs 64 to 66, the copyright policy under Article 53(1)(c) AI Act and the training content summary under Article 53(1)(d) AI Act. This is an opportunity for regulatory coherence that the Guidelines are well placed to seize.

3. Proposals

3.1 New paragraph on the record of the collection context

The EDPB is invited to insert, after paragraph 42 (or as a new item (j) in paragraph 66), the following text:

> *"Controllers should keep, for each collection run, a record of the collection context whose date and integrity can be independently verified, for instance by means of qualified electronic time stamps within the meaning of Article 3(34) of Regulation (EU) No 910/2014 or by tamper-evident, append-only logs. The record should include:*
>
> *(a) the collection criteria, filters and exclusion rules in force (paragraphs 37, 38 and 74);*
> *(b) the identification of the crawler, including the user agent used (paragraph 30);*
> *(c) for each source, the date and time of access and the restrictions observed at that time – such as robots.txt or ai.txt files, reservations of rights within the meaning of Article 4(3) of Directive (EU) 2019/790, authentication requirements or responses that differ from those served to ordinary browsers (paragraphs 37, 64 and 66(c)) – together with the decision taken;*
> *(d) the exclusions and deletions applied, including those under paragraph 74(ii), with the date on which they were applied;*
> *(e) the version of the published list of sources (paragraphs 31 and 66(e)) and of the opt-out list (paragraph 66(f)) applied to the run.*

The record should not contain the personal data collected, but only the information necessary to establish the conditions of the collection; where source identifiers may themselves constitute personal data, they should be recorded at the level of the domain or in pseudonymised form. The record should be retained for as long as the dataset or the model trained on it is used, be made available to the supervisory authority on request and, where the dataset is made available to another entity (paragraph 19), accompany it."

Rationale. A qualified electronic time stamp enjoys a legal presumption of the accuracy of the date and time it indicates and of the integrity of the data to which they are bound (Article 41(2) of Regulation (EU) No 910/2014), and is recognised in all Member States (Article 41(3)). Applied to a daily digest of the crawler's log, it makes the record verifiable at negligible cost; hash-chained, append-only logs offer a technology-neutral alternative. The record is generated by the crawler itself as a by-product of the decisions the Guidelines already require it to take, so the additional burden on developers, including small and open-source developers, is marginal. The exclusion of the collected personal data from the record keeps it consistent with data minimisation and avoids creating a second copy of the dataset.

3.2 Consequential clarifications

- **Paragraph 31.** After the good practice of listing domain names and URLs with the collection date or period, add: "The list should be versioned, so that the information published for each collection period remains available after the list is updated."
- **Paragraph 64, third indent.** Add: "The relevant restrictions are those in force at the time of collection. Since websites change (paragraph 10), the controller should record the restrictions observed at that time (see paragraph [new])."
- **Paragraph 66(e).** Replace "an updated list of scraped websites" by "a versioned list of scraped websites, indicating the collection period for each entry".
- **Paragraph 75.** Add: "This demonstration should rely in particular on the record referred to in paragraph [new], including the versions of the filters applied and the deletion events under paragraph 74(ii)."
- **Footnote to paragraph 66 or to Example 7.** Add a reference to Article 53(1)(c) and (d) of Regulation (EU) 2024/1689, noting that the record of the collection context may serve both the GDPR assessment and the obligations of providers of general-purpose AI models.

3.3 Additional worked example (Section 3.2.3, after Example 7)

In line with the seven examples in the Guidelines, the following example is proposed:

Example 8: Restrictions introduced after collection

An organisation develops a text generation model using content collected from publicly accessible platforms. Its crawler identifies itself with a published user agent, reads the robots.txt and ai.txt files of each site before access, honours reservations of rights expressed under Directive 2019/790, and does not access content behind authentication. For each run, the crawler records the sources accessed, the date and time of access, the restrictions observed and the decision taken, as well as the filters applied and the data excluded. A digest of the record is time-stamped daily by a qualified trust service provider. The organisation publishes a versioned list of sources with the collection period of each entry and maintains an opt-out list.

Eight months after a collection run, a platform amends its robots.txt file to exclude AI crawlers and amends its terms to prohibit the use of user content for AI training. A user of the platform lodges a complaint with the supervisory authority, stating that the platform prohibits scraping. The organisation produces the record. The supervisory authority is able to establish, without relying on the organisation's statements, that at the time of collection the platform displayed no

restriction, that the content was publicly accessible without authentication, that the crawler had identified itself, and that the platform was removed from subsequent runs as soon as the restriction appeared. The supervisory authority also establishes that the complainant's objection was processed under the opt-out list. In those circumstances, the data subject's reasonable expectations are assessed by reference to the situation at the time of collection, and the balancing test may generally be considered as met for that collection.

Conversely, an organisation that cannot show which restrictions were in force at the time of collection cannot rely on Example 4 to establish the reasonable expectations of the data subjects, and the balancing test is carried out on the basis of the restrictions in force at the time of the assessment.

4. Limits of the proposal, stated openly

- The record establishes *what* the crawler observed and did, and *when*. It does not establish that the measures were effective; that remains a matter for the verification and assessment required by paragraph 76 and for the state-of-the-art standard in paragraph 77.
- The record does not replace the public information required under Article 14(5)(b) GDPR (paragraphs 30 to 33); it is the internal counterpart that makes the public information verifiable and allows it to be versioned.
- The proposal does not prejudge the outcome of the balancing test. A record that shows the crawler ignored a restriction is as probative as one that shows it honoured it. This is why a verifiable record deserves weight, and why no controller can use it to manufacture compliance.
- The proposal creates no new obligation. Articles 5(2) and 24 GDPR already require controllers to be able to demonstrate compliance, and Recital 74 extends this to the effectiveness of the measures. The proposal names a proportionate way of doing so for the collection phase.

5. Expected benefits

- Accountability becomes achievable for web scraping, answering the concern expressed in paragraph 14, through a record that the crawler produces as a by-product of the decisions the Guidelines already require.
- Supervisory authorities obtain a common standard for verifying crawl-time facts across Member States, reducing divergent assessments of the same conduct.
- Data subjects obtain the information they need to exercise the rights referred to in paragraphs 60 and 66(f): whether, when and from which source their data was collected, including for past collections.
- Downstream reusers of datasets (paragraphs 19 and 32) receive the collection context they need to carry out their own assessment, so that provenance travels with the data.
- Developers subject to the AI Act can meet the GDPR and the copyright-related obligations of Article 53 with one record rather than two, and the EDPB and the AI Office speak with one voice on the facts that both regimes require.
- Developers that respect restrictions are no longer indistinguishable from those that do not. The incentive created is to build compliance into the crawler, which is data protection by design in the sense of Article 25 GDPR.

I remain at the disposal of the EDPB Secretariat for any clarification on the technical aspects of this contribution, including the implementation of the record described in Section 3.1.

Thiago Francis Alves da Silva
Alpes-Maritimes, France – 28 October 2026