Comments on EDPB Guidelines 03/2026

Technical terminology in paragraphs 7–11

Mag. iur. Michael Fischer | 30 October 2026

1. Scope and purpose

The objectives of Guidelines 03/2026 on web scraping in the context of generative AI, and their general approach to data protection, are broadly supported. However, paragraphs 7 to 11 alternate between the technical terms “web scraping”, “web crawling”, “scraper”, “fetching”, “rendering” and “extraction” without consistently distinguishing the activities to which they refer.[1]

These terms may describe distinct stages of a common automated data-acquisition process. The present comments therefore propose a brief, comprehensible clarification of the terminology and its consistent use. This would neither narrow the scope of the GDPR nor change the data-protection assessment of the activities concerned.

2. Technical distinctions between activities

Web scraping (broad usage). The term may be used as an umbrella expression for automated acquisition of data from web sources. In a narrower technical sense, it refers to selecting and extracting specific information from web content that has already been obtained or made available. In light of these different usages, an express definition for the purposes of the Guidelines would be helpful.[2]

Web crawling. A crawler is an automated client that discovers and retrieves web resources, often by following further links. It may operate broadly, within a restricted set of sources, or with a thematic focus. RFC 9309 addresses such crawlers and rules governing automated access to URI paths through robots.txt; it does not purport to define the entirety of web-data processing.[3]

Fetching or HTTP retrieval. Retrieval involves requesting a resource and transferring a representation, such as HTML or JSON. RFC 9110 describes the GET method in terms of requesting the transfer of a selected representation. The subsequent identification and extraction of particular information from that representation is a technically distinct operation.[4]

Rendering. A browser engine may process HTML and execute client-side JavaScript to make dynamic content or page states available. Such rendering is not necessary for every retrieval operation, nor is it equivalent to selecting particular data fields.[5]

Parsing and data extraction. Parsing makes retrieved content amenable to automated analysis; an extraction component may then select particular elements or data. MDN demonstrates, for example, how HTML may be parsed into a DOM without executing embedded scripts and how CSS selectors may be used to select elements. Technical architectures can likewise separate retrieval (“Fetch”) from analysis (“Parse”).[6]

The case law also distinguishes operations carried out in the course of processing web data: the Court of Justice identifies the automatic finding, retrieval, organisation, storage and making available of information by search

---

[1] *EDPB*, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0 (7 July 2026) paras 7–11, https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf (accessed 30 October 2026).

[2] *Jarmul/Lawson*, Python Web Scraping² (2017) 17, 35; *Smith*, Go Web Scraping Quick Start Guide (2019) 5–6.

[3] *Koster/Illyes/Zeller/Sassman*, Robots Exclusion Protocol, RFC 9309 (2022) sections 1, 2.1, https://www.rfc-editor.org/rfc/rfc9309.html (accessed 30 October 2026); *Olston/Najork*, Web Crawling, Foundations and Trends in Information Retrieval 4/3 (2010) 175 (180–185, 203–213).

[4] *Fielding/Nottingham/Reschke*, HTTP Semantics, RFC 9110 (2022) section 9.3.1, https://www.rfc-editor.org/rfc/rfc9110.html (accessed 30 October 2026); *Mozilla*, GET request method, MDN Web Docs, https://developer.mozilla.org/en-US/docs/Web/HTTP/Reference/Methods/GET (accessed 30 October 2026); *Mozilla*, Using the Fetch API, MDN Web Docs, https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API/Using_Fetch (accessed 30 October 2026).

[5] *Mozilla*, Critical rendering path, MDN Web Docs, https://developer.mozilla.org/en-US/docs/Web/Performance/Guides/Critical_rendering_path (accessed 30 October 2026); *Jarmul/Lawson*, Python Web Scraping² (2017) 104–107.

[6] *Mozilla*, DOMParser: parseFromString() method, MDN Web Docs, https://developer.mozilla.org/en-US/docs/Web/API/DOMParser/parseFromString (accessed 30 October 2026); *Mozilla*, Document: querySelector() method, MDN Web Docs, https://developer.mozilla.org/en-US/docs/Web/API/Document/querySelector (accessed 30 October 2026); *Smith*, Go Web Scraping Quick Start Guide (2019) 105–109.

engines as distinct activities. The judgment does not, however, establish independent technical definitions of crawling or scraping.[7]

3. Points requiring clarification

Paragraph 7. The broad description of "web scraping" is defensible as an umbrella concept. It should, however, be expressly identified as a definition for the purposes of the Guidelines, so that it is not conflated with the narrower operation of data extraction.

Paragraph 8. The EDPB already distinguishes, appropriately, between targeted and untargeted collection by reference to the specificity of the collection criteria rather than the tool employed. Nevertheless, the example of "web crawling without collection criteria" could make clear that crawling can be employed for both targeted and untargeted collection and is not inherently a category of untargeted scraping.[8]

Paragraphs 9 and 10. For static content, the Guidelines already distinguish fetching from parsing. When addressing dynamic content, however, they alternate between "web crawlers" and "scrapers". Describing the function would be clearer: rendering may make dynamic content accessible, while subsequent link discovery and targeted data extraction remain distinguishable activities. Changes to a website's structure may primarily affect extraction rules without preventing HTTP retrieval.

Paragraph 11. The second process step particularly warrants clarification: describing the transfer of data to a storage location as "extraction" primarily captures retrieval and the capture or storage of raw data. Targeted extraction of particular content may take place subsequently, but need not coincide with that operation. It is therefore proposed to distinguish "fetching/retrieval and raw capture" from "content or data extraction".[9]

4. First wording proposal – clarification of technical terms

It is suggested that paragraph 7 expressly adopt a broad umbrella term and that paragraphs 8 to 11 use the terms for individual operations consistently. A concise clarification might read:

> *"For the purposes of these Guidelines, web scraping is used as an umbrella term for the automated acquisition of data from web sources. Depending on the technical design, this may include the selection of sources, the discovery and retrieval of resources (crawling/fetching), the processing of dynamic content (rendering), targeted data extraction, and cleaning, structuring and storage. These stages may be implemented within an integrated system or by separate components."*

This clarification is intended solely to improve technical comprehensibility. Where personal data are processed, the data-protection assessment of each relevant processing operation remains unaffected.[10]

5. Second wording proposal – technology-neutral terminology

The operations described above cover common, but not all conceivable, forms of automated web-data acquisition. For example, information from web sources may also be obtained through application programming interfaces (APIs), data feeds or available web archives, without using a conventional link-following crawler.[11]

Recital 15 GDPR expressly states that the protection of natural persons should be technologically neutral and should not depend on the techniques used. This principle concerns the protection afforded by data-protection law and does not prescribe particular terminology; it nevertheless supports a future-proof technical description. As an alternative, the umbrella expression "automated web-data acquisition" could therefore be adopted to describe present and future methods of acquiring data from web sources without tying the description to any specific technical steps.[12]

The Court of Justice confirms, with reference to recital 15 GDPR, that the protection afforded by data-protection law should not depend on the technology employed.[13] It also emphasises the broad, non-exhaustive definition of

---

[7] CJEU 13 May 2014, C-131/12, *Google Spain and Google*, ECLI:EU:C:2014:317, paras 28, 35-38.

[8] *EDPB*, Guidelines 03/2026 para 8; *Koster/Illyes/Zeller/Sassman*, RFC 9309 section 1.

[9] *EDPB*, Guidelines 03/2026 para 11; *Fielding/Nottingham/Reschke*, RFC 9110 section 9.3.1; *Smith*, Go Web Scraping Quick Start Guide (2019) 105–109.

[10] Art 4(2) GDPR; *Petrlic/Sorge/Ziebarth*, Datenschutz² (2022) 185 f.

[11] Utrecht University, Collecting data from the web, Research Data Management Support, https://www.uu.nl/en/research/research-data-management/guides/during-research/collecting-data-from-the-web (accessed 30 October 2026).

[12] Recital 15 GDPR, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679 (accessed 30 October 2026).

[13] CJEU 7 March 2024, C-740/22, *Endemol Shine Finland*, ECLI:EU:C:2024:216, para 34.

processing in Article 4(2) GDPR.[14] This case law supports future-proof terminology but does not prescribe any particular technical umbrella term.

The following alternative wording is proposed for paragraph 7:

"For the purposes of these Guidelines, automated web-data acquisition means the automated obtaining of information from web sources, irrespective of the particular technical method employed. It may include, in particular, crawling and fetching, browser-based interaction and rendering, retrieval via web APIs, ingestion of data feeds or web archives, as well as parsing and data extraction. This list is non-exhaustive. The data-protection assessment of the processing operations concerned remains unaffected."

This second proposal is intended as a terminological alternative for describing the subject matter of the Guidelines more precisely, not as a proposal to expand their substantive scope under data-protection law.

Declaration on AI-assisted preparation

Generative AI was used to assist with drafting, linguistic refinement and translation of this submission. The substantive arguments and final content remain the responsibility of the submitter.

References

EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, Version 1.0 (7 July 2026), https://www.edpb.europa.eu/system/files/2026-07/edpb_guidelines_2020603_webscraping_v1_en_0.pdf

CJEU, 7 March 2024, C-740/22, Endemol Shine Finland, ECLI:EU:C:2024:216, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62022CJ0740

CJEU, 13 May 2014, C-131/12, Google Spain and Google, ECLI:EU:C:2014:317, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62012CJ0131

CJEU, 22 June 2023, C-579/21, Pankki S, ECLI:EU:C:2023:501, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0579

European Commission, Data protection explained, https://commission.europa.eu/law/law-topic/data-protection/data-protection-explained_en (accessed 30 October 2026)

Fielding/Nottingham/Reschke, HTTP Semantics, RFC 9110 (2022), https://www.rfc-editor.org/rfc/rfc9110.html

Jarmul/Lawson, Python Web Scraping² (Packt 2017)

Koster/Illyes/Zeller/Sassman, Robots Exclusion Protocol, RFC 9309 (2022), https://www.rfc-editor.org/rfc/rfc9309.html

Mozilla (MDN), GET request method, https://developer.mozilla.org/en-US/docs/Web/HTTP/Reference/Methods/GET (accessed 30 October 2026)

Mozilla (MDN), Using the Fetch API, https://developer.mozilla.org/en-US/docs/Web/API/Fetch_API/Using_Fetch (accessed 30 October 2026)

Mozilla (MDN), DOMParser: parseFromString() method, https://developer.mozilla.org/en-US/docs/Web/API/DOMParser/parseFromString (accessed 30 October 2026)

Mozilla (MDN), Document: querySelector() method, https://developer.mozilla.org/en-US/docs/Web/API/Document/querySelector (accessed 30 October 2026)

Mozilla (MDN), Critical rendering path, https://developer.mozilla.org/en-US/docs/Web/Performance/Guides/Critical_rendering_path (accessed 30 October 2026)

Olston/Najork, Web Crawling, Foundations and Trends in Information Retrieval 4/3 (2010) 175–246, DOI: 10.1561/1500000017

Petrlic/Sorge/Ziebarth, Datenschutz. Einführung in technischen Datenschutz, Datenschutzrecht und angewandte Kryptographie² (2022)

Smith, Go Web Scraping Quick Start Guide (Packt 2019)

Utrecht University, Collecting data from the web, Research Data Management Support, https://www.uu.nl/en/research/research-data-management/guides/during-research/collecting-data-from-the-web (accessed 30 October 2026)

[14] CJEU 22 June 2023, C-579/21, *Pankki S*, ECLI:EU:C:2023:501, para 46.