CONSULTATIVE REMARKS

ON THE DRAFT EDPB GUIDELINES 03/2026 ON WEB SCRAPING IN THE CONTEXT OF GENERATIVE AI

DR. ZSOLT G. BALOGH

LUDOVIKA UNIVERSITY OF BUDAPEST

INSTITUTE OF THE INFORMATION SOCIETY

UNESCO CHAIR

29 OCTOBER 2026

WEB SCRAPING FOR GENERATIVE AI: TOWARDS EFFECTIVE AND COHERENT DATA PROTECTION

1 INTRODUCTION: FROM PUBLIC ACCESSIBILITY TO ACCOUNTABLE REUSE

1.1 GENERAL ASSESSMENT

The EDPB Guidelines provide a welcome clarification that the public accessibility of personal data does not remove its processing from the GDPR. Their practical significance lies in addressing the transformation of dispersed online information into datasets supporting generative AI. This transformation changes the scale, context, persistence, and potential consequences of processing. Accordingly, the central regulatory question concerns whether a particular collection and subsequent use can be justified, rather than whether information was technically accessible.

The Guidelines should nevertheless distinguish more consistently between binding requirements, interpretative positions, and recommended safeguards. Controllers and processors need predictable compliance expectations; data subjects require enforceable protections; national supervisory authorities need sufficiently precise assessment criteria. Repeated references to case-by-case evaluation are appropriate, but should be supported by common evidential standards.

1.2 SCOPE AND CONCEPTUAL COHERENCE

Excluding public-sector scraping, internal datasets, and certain data-broker activities narrows the Guidelines' usefulness. Although these exclusions do not imply exemption from the GDPR, they leave significant interfaces insufficiently explained. Dataset acquisition may separate collection from training across several organisations and jurisdictions. The final Guidelines

should therefore explain how their conclusions apply at those interfaces, particularly where developers acquire datasets without reliable provenance.

2 DEFINING WEB SCRAPING AND ALLOCATING RESPONSIBILITY

2.1 A FUNCTIONAL DEFINITION

The distinction between targeted and untargeted scraping is helpful, but targeting does not necessarily reduce interference. A narrowly focused collection of health discussions may create greater risks than a broader collection of impersonal technical material. Classification should therefore incorporate sensitivity, identifiability, vulnerability, collection frequency, and intended use, alongside technical breadth.

The literature demonstrates that scraping raises overlapping questions of privacy, intellectual property, contractual obligations, competition, and access to information. Pagallo and Ciani Sciolla's analysis particularly supports assessing purposes and consequences rather than treating scraping as an inherently homogeneous activity.[1] The Guidelines should also clarify their boundaries concerning API extraction, purchased web archives, and retrieval used during inference, even where these practices fall outside their immediate training-focused scope.

2.2 RESPONSIBILITY ACROSS THE SUPPLY CHAIN

Paragraphs 16–20 correctly adopt a factual approach to controllership. However, the distinction between independently collected datasets and commissioned scraping requires stronger operational guidance. Commercial descriptions such as "data supplier" cannot determine GDPR status. A scraper selecting sources for its own resale purposes will ordinarily act as a controller, whereas an entity operating exclusively under documented instructions may qualify as a processor.

Developers should undertake proportionate provenance checks covering collection methods, lawful bases, restrictions, sensitive-data filtering, and rights-management arrangements. Such checks do not automatically establish joint controllership, but support accountability for the

[1] Pagallo, U. and Sciolla C. J. (2023) pp. 2, 10

developer's own processing. Contractual assurances should be corroborated where identifiable risks make reliance on them unreasonable.[2]

3 APPLYING THE GDPR WITHOUT DILUTING ITS GUARANTEES

3.1 LEGITIMATE INTERESTS AND REASONABLE EXPECTATIONS

The Guidelines appropriately require separate assessments of legitimate interest, necessity, and balancing. Nevertheless, "developing generative AI" is too general to demonstrate necessity. Controllers should identify intended capabilities, explain the contribution of personal data, and assess less intrusive alternatives.

Example 4 deserves particular revision. A platform's announcement that scraping may occur cannot establish that users reasonably expect unrestricted third-party AI training. Awareness of a practice differs from its normative acceptability. Kokoulina's analysis exposes the instability of treating public accessibility[3] and reasonable expectations as straightforward technical facts, while Solove and Hartzog demonstrate the wider conflict between indiscriminate scraping and privacy principles.[4]

Likewise, the absence of robots.txt restrictions should not favour processing as though individuals had waived protection. Many data subjects cannot control website architecture. Example 7 should avoid suggesting that a collection of safeguards generally establishes a favourable balance without examining the actual interests and consequences involved.

3.2 PURPOSE LIMITATION, MINIMISATION, AND TRANSPARENCY

Purpose specification must be sufficiently concrete to constrain collection and subsequent reuse. Where further processing is undertaken by the same controller, Article 6(4) may require a compatibility assessment; a separate recipient must establish the lawfulness of its own processing. Neither public availability nor a general innovation objective resolves these questions.

---

[2] GDPR, Arts. 5(2), 24, 26, and 28

[3] Kokoulina, O. (2024) pp. 715, 718

[4] Solove and Hartzog (2025) pp. 1554-1555

Minimisation should be demonstrated through collection criteria, comparative dataset testing, retention limits, and documented exclusions. In *Schrems v. Meta Platforms Ireland*, the Court rejected unrestricted aggregation for advertising purposes. Although that judgment does not determine AI-training legality, its reasoning challenges assumptions that potentially useful information may be accumulated without meaningful limits.[5]

The treatment of Article 14(5)(b) should similarly resist turning operational scale into a routine transparency exemption. Controllers invoking disproportionate effort should document their assessment and adopt effective alternative communication. Searchable source information, identifiable crawler notices, and accessible rights portals would often offer greater practical value than a general privacy policy.

3.3 SENSITIVE DATA AND THE SEARCH-ENGINE ANALOGY

The proposed extension of *GC and Others* to incidental sensitive-data collection is the Guidelines' most contestable interpretation. That judgment concerned search-engine indexing and dereferencing; generative training involves a different transformation of information and potentially different possibilities for extraction and dissemination. Functional similarity requires demonstration rather than assumption. [6]

The EDPB should explain precisely whether its approach concerns the extent of responsibility, enforcement expectations, or substantive lawfulness. An inability to eliminate sensitive data cannot itself create an Article 9(2) derogation. Output filtering also cannot retrospectively establish the lawfulness of collection. Any reliance on the analogy should require evidenced preventive measures, prompt removal, independent evaluation, and explicit consideration of Article 10 data.

3.4 EFFECTIVE RIGHTS AND ANONYMISATION

Data subjects need practical mechanisms to discover processing and seek access, objection, rectification, or erasure. These mechanisms should distinguish source records, training datasets, and personal data processed through model operation. Deleting a dataset entry does not necessarily eliminate information retained by a model.

---

[5] GDPR, Arts. 5(1)(b)–(c), 6(4), 14(5)(b), and 25; C-446/21, paras. 47–65.

[6] C-136/17, especially paras. 45–48. The criticism concerns the proposed extension of the judgment, rather than denying the judgment's relevance.

Pseudonymisation remains a safeguard within the GDPR. Anonymisation requires a separate, context-sensitive assessment; replacing names or producing synthetic data does not automatically achieve it. Controllers should assess memorisation, extraction, and reidentification risks. Article 11 should prevent unnecessary identification without becoming a general justification for designing systems that frustrate rights.[7]

4 COHERENCE WITH THE WIDER DIGITAL REGULATORY FRAMEWORK

4.1 AI ACT AND COPYRIGHT

The AI Act complements GDPR accountability without supplying a general authorisation to scrape personal data. Article 53(1)(c)–(d) requires relevant general-purpose AI providers to maintain copyright-compliance policies and publish training-content summaries. These summaries may assist scrutiny but do not replace Article 14 information or individual rights.[8]

The final Guidelines should also reflect legislative developments after their adoption. The amended AI Act's Article 4a addresses exceptional processing of special-category data for bias detection and correction under defined safeguards; it should not be interpreted as permission for indiscriminate training-data collection.[9]

Hamann's examination of machine-readable reservations demonstrates why technical signals require careful legal interpretation[10]. A copyright reservation under Article 4(3) of Directive 2019/790, a website access restriction, and an individual's GDPR objection have different holders, conditions, and effects. Compliance with one cannot substitute for compliance with the others.[11]

---

[7] GDPR, recital 26 and Arts. 11, 15–19, and 21

[8] AI Act. Art. 53(1)(c)–(d)

[9] AI Act Art. 4a, as inserted by Digital Omnibus on AI. This amendment postdates the consultation version assessed here.

[10] Hamann (2024), p. 110

[11] DSM Directive, Arts. 3–4. Copyright reservations and GDPR objections should remain analytically separate.

4.2 CYBERSECURITY OBLIGATIONS

The Guidelines should connect GDPR Articles 25 and 32 with cybersecurity risk management. Scraping can expose authentication material, concentrate sensitive information, or impair source services. Applicable safeguards include access controls, collection-rate limits, dataset isolation, provenance logging, and incident procedures.

NIS2 applies only where the relevant entity falls within its scope under applicable implementing legislation.[12] Its risk-management and significant-incident reporting requirements must be distinguished from GDPR personal-data-breach obligations: the relevant thresholds and procedures are not interchangeable.[13]

4.3 PLATFORMS AND DATA-ACCESS LEGISLATION

The DMA and DSA establish specific access arrangements rather than a general entitlement to scrape. DMA Article 6(11) concerns access to specified search data, including anonymisation requirements. DSA Article 40 supports defined research into systemic risks; it does not authorise unrestricted commercial model training.[14]

Similarly, the Data Act's access framework, particularly for connected-product and related-service data, does not create a general right to harvest websites. The Data Governance Act facilitates specified reuse and governance arrangements without independently supplying a GDPR legal basis for personal-data processing.[15] The Guidelines should explain these boundaries while recognising that privacy arguments can also be invoked strategically to obstruct legitimate scrutiny or competition.

5 SUPERVISORY PRACTICE AND PROPOSED DIRECTION

5.1 EVIDENCE, REMEDIES, AND COORDINATION

National supervisory authorities should assess demonstrable compliance rather than accept generic descriptions of safeguards. Relevant evidence includes provenance records, necessity

[12] NIS2, Arts. 2–3 and 21–23; NIS2 Implementing Regulation, where applicable.

[13] GDPR, Arts. 25 and 32–34

[14] DMA, Art. 6(11); DSA Art. 40, particularly paras. 4, 8, and 12

[15] Data Act, Art. 1(5) and Chapter II; Data Governance Act, Art. 1(3)

assessments, filtering evaluations, Article 14 exemption analyses, and tests of rights-handling procedures.

A harmonised supervisory methodology would improve predictability while preserving context-sensitive decisions. Authorities should distinguish defects that can be remedied prospectively from unlawful processing requiring cessation, deletion, or further corrective measures. Coordination with AI, cybersecurity, competition, and platform regulators should respect their distinct mandates.

5.2 CONCLUDING POSITION

The Guidelines should be strengthened through clearer supply-chain responsibilities, a more demanding account of reasonable expectations, and a narrower, better justified treatment of incidental sensitive-data collection. Their success should be measured by whether controllers can demonstrate lawful processing, individuals can exercise meaningful rights, and supervisory authorities can verify both.

REFERENCES

ACADEMIC LITERATURE

- Hamann, H. (2024). Artificial intelligence and the law of machine-readability: A review of human-to-machine communication protocols and their (in)compatibility with Article 4(3) of the Copyright DSM Directive. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law, 15*(2), 102–121.
- Kokoulina, O. (2024). Towards future-proof, rights-respecting automated data collection: An examination of European jurisprudence. *Vanderbilt Journal of Entertainment & Technology Law, 26*(4), 707–770.
- Pagallo, U., & Sciolla, C. J. (2023). Anatomy of web data scraping: Ethics, standards, and the troubles of the law. *European Journal of Privacy Law & Technologies, 2023*(2), 1–19.
- Sobel, B. L. W. (2021). A new common law of web scraping. *Lewis & Clark Law Review, 25*(1), 147–208.
- Solove, D. J., & Hartzog, W. (2025). The great scrape: The clash between scraping and privacy. *California Law Review, 113*(5), 1521–1584. https://doi.org/10.15779/Z38H98ZF9D

GUIDELINES AND OPINIONS

- European Data Protection Board. (2024, December 17). *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*.

CASES

- C-446/21, ECLI:EU:C:2024:834 (2024). *Schrems v. Meta Platforms Ireland Ltd*,
- C-136/17 (2019). *GC and Others v. Commission nationale de l'informatique et des libertés*,

LEGISLATION

- 2016 (General Data Protection Regulation)
- Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 (protection of personal data by Union institutions, bodies, offices and agencies)
- DSM Directive. Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market
- Data Governance Act. Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022

- DMA. Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 (Digital Markets Act)
- DSA. Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 (Digital Services Act)
- NIS2 Directive. Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 (NIS2 Directive)
- Data Act. Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 (Data Act)
- AI Act. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (Artificial Intelligence Act)
- NIS2 Implementing Regulation. Commission Implementing Regulation 2024/2690 of 17 October 2024 laying down rules for the application of Directive (EU) 2022/2555
- CRA. Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 (Cyber Resilience Act)
- Digital Omnibus on AI. Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)