October 2026

Fraunhofer

Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e. V.

Comments on the EDPB Guidelines 03/2026 on web scraping in the context of generative AI

A clear and consistent data protection framework for the training of generative AI is essential to strengthen Europe's position in international AI competition while respecting the fundamental rights of individuals. A viable approach can only succeed if the obligations it imposes are technically achievable, consistent with the AI Act and the General Data Protection Regulation (GDPR), and applicable to all actors developing AI models in Europe – including publicly funded research organisations, SMEs and open-source initiatives.

Fraunhofer welcomes the EDPB's initiative to clarify the application of the GDPR to web scraping for generative AI. Where supervisory authorities, research actors and AI developers align on shared standards and technically realistic expectations, a balanced ecosystem becomes possible: research institutions and developers can train models with legal certainty, data subjects can exercise their rights effectively, and trustworthy applications can flourish within clear boundaries. Legal certainty regarding the lawfulness of data collection is a precondition for Europe's ability to build competitive and trustworthy models domestically – and thus for European technological sovereignty.

At the same time, several passages risk creating obligations that are technically unachievable, that duplicate or contradict the AI Act, or that would in practice prevent the development of general-purpose models by European public research organisations, SMEs and open-source initiatives.

Scope: Scientific research and open-source development must be addressed explicitly

The guidelines expressly limit themselves to "web scraping done by private entities" (paras. 6, 43), on the assumption that scraping for generative AI is "mainly performed by private sector bodies". This assumption no longer reflects reality: publicly funded research organisations scrape data for generative AI training purposes or use scraped data collected by private or publicly funded entities for generative AI training. The consequence is a gap in legal certainty precisely for those actors the EU seeks to strengthen, and a skewed legal-basis analysis, since the research-specific provisions of the GDPR are left unaddressed.

This gap is difficult to reconcile with the status the GDPR itself accords to scientific research. Conducting scientific research is recognised as a fundamental freedom under EU law, and Art. 5(1)(b) GDPR establishes a statutory compatibility presumption under which further processing of data originally collected for another purpose is not considered incompatible with the initial purpose where it is carried out for scientific research purposes within the safeguards of Art. 89(1) GDPR. Guidelines that address only private-sector scraping as a primary collection activity fail to engage precisely with the circumstances in which this presumption is most relevant: research organisations and their commercial or open-source partners reusing previously scraped data for generative AI research. For the avoidance of doubt, the compatibility presumption under Art. 5(1)(b) GDPR does not in itself constitute a legal basis for processing and operates in

conjunction with the requirements of Art.6 GDPR and, where applicable, Art. 9 GDPR.

The final guidelines should therefore address the interaction with Art. 5(1)(b), Art. 5(1)(e), Art. 14(5)(b) and Art. 89(1) GDPR for scraping carried out for scientific research purposes – including where research results are subsequently exploited commercially or released under permissive open-source licenses that allow commercial use, such as Apache 2.0. They should further confirm that scientific research may constitute a legitimate interest within the meaning of Art. 6(1)(f) GDPR, where that provision is applicable to the controller concerned, complementing the examples in para. 50, and that the weight of that interest in the balancing test reflects the public interest in research recognised in recital 159 GDPR and Art. 13 of the Charter of Fundamental Rights, consistent with the EDPB's own Guidelines 1/2026 on the processing of personal data for scientific research purposes, which confirm that scientific research can constitute a legitimate interest regardless of its commercial or non-commercial nature and will often carry significant weight in the balancing test. Example 2 of those Guidelines 1/2026 – a start-up conducting research on bias in generative AI models in partnership with a university faculty, whose results are published and may also be exploited commercially – illustrates that the EDPB itself already recognises exactly this type of actor and research activity as falling within the scope of scientific research under the GDPR; the web scraping guidelines should adopt a consistent approach.

Roles: Clear Allocation of Responsibility Along the Data Supply Chain

Fraunhofer strongly welcomes para. 19 and its confirmation that the scraper and the AI developer are, in principle, responsible for their own separate processing activities. This clarification is of central practical importance for the use of large public web corpora such as Common Crawl and derived datasets, on which European open model development substantially depends.

The guidelines should, however, draw out the consequences. The re-user's obligations should be assessed in relation to the processing it actually controls, as collection decisions taken by a third party cannot be changed retroactively, without prejudice to its responsibility for the lawfulness of its own processing. Where a developer obtains an existing corpus, it does not necessarily have access to the original crawler configuration, to the robots.txt state at the time of crawling, or to URL-level provenance. Widely used and well-documented public corpora should therefore be recognised as a distinct category, accompanied by guidance on what constitutes adequate due diligence, for example reliance on dataset documentation and datasheets, verification that the provider honours removal requests, and the application of state-of-the-art PII filtering.

This distinction must also be reflected in the data minimisation section. Para. 37 lists measures the controller should take *before* the collection of the data. Where an AI developer uses an existing crawled dataset, none of these ex ante measures is available to it. Fraunhofer recommends restructuring paras. 37–38 to distinguish explicitly between a controller performing its own crawl, for which the ex ante measures apply, and a controller re-using an existing corpus, whose obligations – consistent with the necessity and data minimisation requirements under Art. 6(1)(f) GDPR as interpreted by the Court of Justice – may depending on the circumstances, require conducting due diligence on the dataset and applying its own filtering at the point of use.

Transparency: Align with the EU AI Act Rather Than Create a Parallel Regime

For the training of general-purpose AI models, transparency expectations should align with those established in the AI Act, while recognising that the respective transparency obligations under the GDPR and the AI Act pursue distinct legal purposes without creating a parallel and more granular regime. The AI Act's framework – requiring providers to draw up and make publicly available a sufficiently detailed summary of the content used for training – represents a reasonable and workable reference point. Disclosure should accordingly focus on meaningful information such as categories and types of sources, named base datasets and versions, crawler characteristics, collection periods and filtering methodologies, rather than on exhaustive inventories of domains and URLs.

Security: Establish Standards for the Data Supply Chain

Compliance with the principles and requirements of the GDPR should be ensured through appropriate measures at all stages of the data supply chain. Establishing standards at each step of the chain would also include specifying evaluation measures, which should be based on existing evaluation measures from the literature which have been accepted by the scientific community. In particular, at the data collection stage, the party responsible for the data collection via web scraping should comply with the requirements of the GDPR regarding safe storage of corpora which may contain personal data. This includes protection against unauthorised or unlawful processing appropriate to the level of risk (Art. 5(1)(f) and Art. 32 GDPR). PII prevalence should be evaluated, for example using exact count of mail addresses, phone numbers, and IP addresses based on regular expression matches (Subramani et al., 2023; Elazar et al, 2024) or on the output of Named-Entity Recognition systems (Hong et al., 2026). Similarly to language technologies being dual-use, also web-scraped corpora may have dual-use characteristics, as they can be maliciously used for harmful and privacy-infringing purposes.

At the distribution stage, data licenses or data-use agreements can provide a helpful tool to establish contractual restrictions on access and downstream use of web-scraped data, by defining permitted purposes and prohibiting harmful use. Lastly, at the model training and evaluation stage, the AI Act requires that GPAI models with systemic risks trained on such data are evaluated to mitigate systemic risks (Art 55(1)(a)–(b) AI Act). The guidelines should also specify that model evaluation should assess whether personal data contained in the training corpora was memorised and could subsequently be extracted or otherwise reproduced by prompting the model. This is particularly relevant if as a result of the balancing exercise the removal of PII is deemed unfeasible. For research organisations in particular, the guidelines should recognise pseudonymisation, together with other appropriate technical and organisational measure, as an appropriate safeguard within the meaning of Art. 89(1) GDPR where complete removal of personal data from large, heterogeneous web corpora is not achievable with state-of-the-art filtering without preventing or seriously impairing the achievement of the research purposes: pseudonymization reduces risk to data subjects while preserving the scientific and technical usability of the dataset, and any residual risk should then be mitigated through appropriate safeguards and taken into account in the balancing exercise

and the memorisation safeguards described above, rather than through an unattainable completeness standard for PII removal.

Extensive research work has measured eidetic memorization of PII in LLMs (Carlini et al., 2021; Carlini et al., 2023; Nasr et al., 2023) - namely, a model's tendency to output training examples verbatim, including PII - by sampling large numbers of generations, ranking them with membership-inference scores, and checking them against web searches and the original training data; these studies suggest that LLMs should be audited for memorisation using such methods.

Worked Examples Should Not Generalise Into New Rules

Example 3 should not become a de facto notification threshold. The example concludes that a controller must notify a closed group of 5,000 identifiable individuals individually. Read in isolation, the figure of 5,000 may be understood as a numerical benchmark for determining whether individual notification is required, which is neither the intended reading nor a consequence of Art. 14(5)(b). The guidelines should state expressly that the assessment is qualitative and that no numerical threshold is intended and should add a counter-example involving a comparable number of data subjects where notification is *not* required, so that the genuinely distinguishing factors – the intrusiveness of contacting scraped addresses, the reliability of the contact details, the effort involved and the availability of public notification – become visible.

Example 7 should not import copyright compliance into the balancing test. The example lists, among the factors making the balancing test "generally … considered as met", that the controller "excludes any content protected by copyright". Compliance with Directive (EU) 2019/790 should remain distinct from the balancing exercise under Art. 6(1)(f) GDPR the Art. 6(1)(f) assessment: copyright protects rights holders' economic interests, whereas the balancing test concerns the fundamental rights of data subjects, as confirmed by the EDPB's own Guidelines 1/2024 on Article 6(1)(f) GDPR, which limit the balancing exercise to the data subjects' own interests, fundamental rights and freedoms. Making one a proxy for the other creates cross-regime conditionality that neither instrument foresees, nor suggests that lawful reliance on legitimate interest depends on training exclusively on public-domain content. Fraunhofer recommends clarifying that compliance with other legal regimes is required independently of and should not by itself determine the outcome of the Art. 6(1)(f) assessment.

Special Categories of Data

Fraunhofer welcomes the application of the "responsibilities, powers and capabilities" test (paras. 72–74) to incidental and residual processing, which addresses a genuine and otherwise intractable problem. This test is derived by the EDPB from the Court of Justice's judgment in GC and Others (De-referencing of sensitive data), C-136/17, paras. 37 and 44–47, which ties the search engine operator's responsibility for special categories of data specifically to the referencing of a web page and the display of a link in a name-based search result. Condition (a) – that the processing has "relevant similarities with the processing activity of a search engine" – relies on an insufficiently explained and potentially narrow analogy to that judgment. Web scraping for AI training, and the subsequent training of a generative AI model, differ materially in important respects from the referencing and ranking of web pages by a search engine.

This creates an unresolved tension: the EDPB considers the Court's ruling relevant to web scraping for AI training, while limiting that analogy to activities similar to those examined in the judgment and leaving that assessment to controllers. The guidelines neither define the required similarities with sufficient precision nor explain how web scraping for model training could satisfy this condition. In GC and Others, the Court emphasised that name-based search results provide internet users with a structured overview enabling them to establish a detailed profile of the data subject (C-136/17, paras. 36–37) – an element not necessarily present in the same form where data is ingested into a persistent training corpus without any query-based retrieval of personal data. Live web searches or the retrieval stage of a RAG system may closely resemble search-engine processing. By contrast, scraping websites to create a persistent training corpus bears a materially different and potentially weaker resemblance to the processing examined by the Court. The guidelines should therefore identify the decisive characteristics and distinguish between search or retrieval, persistent data ingestion, and model training.

References

Elazar, Y., et al. (2024). What's in my big data? International Conference on Learning Representations (ICLR 2024).

Subramani, N., Luccioni, S., Dodge, J., & Mitchell, M. (2023). Detecting personal information in training corpora: An analysis. Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), 208–220.

Hong, R., Hutson, J., Agnew, W., Huda, I., Kohno, T., & Morgenstern, J. (2026). A common pool of privacy problems: Legal and technical lessons from a large-scale web-scraped machine learning dataset. Proceedings of the Symposium on Computer Science and Law, 1–14.

Carlini, N., et al. (2021). Extracting training data from large language models. Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), 2633–2650.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Quantifying memorization across neural language models. International Conference on Learning Representations (ICLR 2023).

Nasr, M., et al. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.