JurisPONt
Analysis

Web Scraping for Generative AI: What the EDPB’s Draft Guidelines Change

Editorial Team

29 October 2026 · EN

Abstract

The EDPB’s draft Guidelines 03/2026 bring web scraping for generative AI firmly within the GDPR compliance debate. This analysis examines public accessibility, purpose limitation, legitimate interest, data minimisation, accuracy, transparency, sensitive data and the growing importance of dataset provenance in AI governance.

The enormous appetite of generative AI for data has turned web scraping from a largely technical practice into a legal question of first importance. Public websites, professional profiles, social networks, forums and image repositories offer material at a scale that would have been unimaginable only a few years ago. For developers, automated collection is efficient. For data protection law, however, efficiency answers very little.

On 8 July 2026, the European Data Protection Board adopted Guidelines 03/2026 on web scraping in the context of generative AI for public consultation. The consultation remains open until 30 October 2026. The text is therefore not yet final, but its significance should not be underestimated. It gives a fairly clear indication of how European supervisory authorities expect the GDPR to operate when personal data available online are collected on a large scale for AI development.

The Board does not attempt to outlaw scraping. Its approach is more demanding than that. It asks controllers to justify the chain of processing from the moment data are collected to their eventual use in an AI model.

Public access is not a legal basis

The first point is also the most easily misunderstood. Information being visible on the internet does not place it outside the GDPR.

Where scraping involves the collection, storage, organisation or retrieval of personal data, those operations constitute processing. The familiar argument that “the information was already public” therefore cannot, by itself, answer the question of lawfulness.

That distinction matters particularly in AI development. A person may publish a professional biography because it is useful to prospective clients, contribute to a specialist forum in order to participate in a particular discussion, or upload an image for a limited social purpose. None of those acts necessarily carries an expectation that the material will later be harvested in bulk and incorporated into a training dataset.

Public availability is relevant to the legal assessment. It is not the end of that assessment.

Scraping forces controllers to define what they are actually doing

One of the strongest aspects of the EDPB's approach is its insistence on purpose limitation.

Large-scale data acquisition can easily obscure the distinction between having the technical capacity to collect information and having a legally defensible reason to do so. A purpose stated simply as "training AI" or "improving the model" risks being too broad to explain why particular personal data are necessary.

This is not a merely formal problem. The purpose determines much of what follows: which sources are appropriate, which categories of information can be collected, how long data should be retained, and whether the interests of the controller can justify the effects on the individuals concerned.

Web scraping brings together information created in entirely different social and professional contexts. The resulting dataset may erase those distinctions. Data protection law, by contrast, requires controllers to take them seriously.

The tension is therefore not between innovation and privacy in the abstract. It lies between the technological logic of aggregation and a legal framework that demands specificity.

Legitimate interest remains available — but it is not a shortcut

For many AI developers, the most consequential part of the Guidelines will be the discussion of legitimate interest under Article 6(1)(f) GDPR.

The EDPB does not say that legitimate interest is unavailable for AI training based on scraped data. That would have been a far more restrictive position. Instead, the Board builds on its Opinion 28/2024 on AI models and insists on the ordinary three-stage analysis.

There must first be a legitimate interest that is lawful, sufficiently precise and genuinely present rather than speculative. The processing must then be necessary to pursue that interest. Finally, the controller must balance the interest pursued against the interests, rights and freedoms of the people whose data are involved.

In web scraping, the balancing stage is likely to do much of the work.

Whether information was publicly accessible is relevant, but so is the context in which it appeared. A controller may need to consider the nature of the website, the sensitivity of the material, the relationship — if any — between the individual and the organisation collecting the data, the privacy settings available, the potential uses of the resulting model and, importantly, what the person could reasonably have expected.

This makes legitimate interest a possible basis for processing, not a general licence for data acquisition.

The practical consequence is that two scraping operations aimed at broadly similar AI systems may produce different legal outcomes because their sources, safeguards and uses differ.

Data minimisation challenges the “more is better” instinct

Generative AI has often developed under an engineering assumption that larger datasets are desirable. The GDPR starts from a different premise: personal data should be adequate, relevant and limited to what is necessary.

Those two logics do not necessarily make AI training impossible, but they force developers to make choices.

The EDPB expects controllers to consider whether particular categories of information can be excluded and whether some sources should not be scraped at all. Its earlier Opinion on AI models already pointed to safeguards such as excluding especially sensitive sources and respecting mechanisms through which websites object to automated collection.

This shifts data protection upstream.

A company cannot simply assemble the largest dataset it can obtain and ask compliance teams to deal with the consequences afterwards. Decisions about source selection, filters and exclusion criteria increasingly form part of the legal design of the system itself.

That is perhaps one of the broader lessons of the Guidelines: data governance begins before model training.

Accuracy is not only a problem of model output

Debates about generative AI frequently focus on hallucinations and inaccurate outputs. The EDPB brings attention to an earlier stage.

Material collected online may already be wrong, obsolete, duplicated or stripped of the context that originally gave it meaning. When such material enters a very large training corpus, correcting it later may be difficult or impossible in practice.

The Board therefore recommends collecting from reliable sources, recording when information was obtained and validating data before using them for AI training.

This has an important implication for AI governance. Dataset provenance is not merely a matter of good engineering practice. It is connected to the GDPR principle of accuracy.

The quality of a model’s informational foundations can therefore become part of the controller’s legal accountability.

Transparency cannot disappear simply because the dataset is enormous

Web scraping is usually invisible to the people whose information is collected. At scale, individual notification may also be exceptionally difficult.

The EDPB acknowledges that the GDPR itself provides for circumstances in which providing information individually may be impossible or involve disproportionate effort. But this cannot be converted into a general transparency exemption for AI development.

That distinction is important.

A controller still has to analyse which information obligations apply and whether relying on an exception is justified in the particular circumstances. Where individual notice is not required, other forms of transparency may become more important: publicly accessible information about scraping practices, intelligible explanations of purposes and meaningful mechanisms through which people can understand or challenge the use of their data.

The scale of the processing may affect how transparency is delivered. It does not remove the underlying obligation to account for it.

Sensitive data expose the limits of indiscriminate collection

The problem becomes sharper when scraped datasets contain special categories of personal data.

Political opinions, religious beliefs, health information, racial or ethnic origin and sexual orientation may all appear in publicly available material. Their presence may be intentional, incidental or difficult to detect before collection.

Article 9 GDPR nevertheless remains applicable.

The EDPB recalls that controllers dealing with such information need both a lawful basis under Article 6 and an applicable condition under Article 9(2). The existence of publicly accessible sensitive information does not create a general exemption.

The Board refers in this context to the Court of Justice's judgment in *GC and Others* (C-136/17), while treating its reasoning cautiously. The judgment may matter where special-category data are captured incidentally or residually, but it does not remove the obligation to assess the particular circumstances and to implement appropriate technical and organisational safeguards.

For large-scale AI datasets, that may require filtering at or close to the point of collection rather than relying solely on downstream controls.

The real issue is becoming provenance

The Guidelines are formally concerned with scraping, but their implications extend into the later life of the model.

That is where the EDPB's 2024 Opinion on AI models becomes particularly important. If personal data were processed unlawfully during development, that history may affect the legal assessment of subsequent processing involving the model. The answer depends on

the circumstances, including whether the model can genuinely be regarded as anonymous.

This makes provenance more than a technical record of where data came from.

An AI provider may eventually need to explain which sources were used, which legal basis supported the collection, what exclusions were applied, how sensitive material was addressed and what steps were taken to protect the rights of individuals.

A model may be deployed months or years after the underlying scraping occurred. The legal relevance of the collection phase does not necessarily end when training does.

For companies developing generative AI, maintaining a defensible record of the data pipeline is therefore becoming part of regulatory risk management.

From permission to accountability

The most useful way to read Guidelines 03/2026 is not as an attempt to answer the simplistic question “Is web scraping legal?”

There is unlikely to be one answer.

What emerges instead is a model of accountable scraping. The important question is whether a controller can demonstrate why a particular operation, involving particular sources and particular categories of data, is compatible with the GDPR.

That moves the debate away from a binary opposition between open internet and data protection.

The internet may make information technically accessible. The GDPR asks a different set of questions: why is it being collected, is that collection necessary, what could the person reasonably have expected, what risks follow from the reuse, and what safeguards have been put in place?

For AI developers, those questions are becoming part of product design rather than an afterthought.

Conclusion

The EDPB’s draft Guidelines do not close the door on web scraping for generative AI. They do, however, make clear that scale and public accessibility cannot replace legal justification.

A developer relying on online personal data will increasingly need to account for the origin of those data, the purpose of their collection, the legal basis relied upon, the expectations of the people concerned and the safeguards built into the acquisition process.

The final Guidelines may change following consultation. But the direction is already visible.

European data protection law is moving the scrutiny of AI further upstream — from the output of the model to the choices that shaped its training data. In that environment, the question for an AI provider will no longer be only whether data can be scraped. It will be whether the provider can defend the way they were obtained.

Sources

European Data Protection Board, *Guidelines 03/2026 on web scraping in the context of generative AI*, version adopted for public consultation, 8 July 2026.

European Data Protection Board, *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*, 18 December 2024.

Court of Justice of the European Union, *GC and Others v CNIL*, C-136/17, judgment of 24 September 2019, ECLI:EU:C:2019:773.

Regulation (EU) 2016/679 (General Data Protection Regulation), in particular Articles 5, 6, 9, 13 and 14.

How to cite

Editorial Team, ‘Web Scraping for Generative AI: What the EDPB’s Draft Guidelines Change’ (JurisPONt, 29 October 2026) <https://jurispont.com/en/articles/web-scraping-generative-ai-edpb-guidelines-2026>