HOGAN LOVELLS
CADWALADER

News

EDPB publishes guidelines on web scraping for generative AI training

22 July 2026

On 7 July 2026, the European Data Protection Board ("EDPB") adopted Guidelines 03/2026 on web scraping in the context of generative AI (version 1.0) ("Guidelines"), now open for public consultation. The Guidelines address the GDPR implications of using automated tools to extract personal data from publicly available web sources for the purpose of developing and training generative AI models. They cover situations where a private entity scrapes data from internet sources external to the organization, either directly or by contracting another party, or obtains already scraped datasets from a data broker. This article summarizes the key findings.

I. Background

Web scraping is a technique that uses automated tools for extracting and storing information from publicly available web services, such as public registers, open-data portals, news outlets, social media, discussion forums, and personal websites. Many of these websites are likely to contain personal data.

In its draft Guidelines 03/2026 on web scraping in the context of generative AI, the EDPB distinguishes between *targeted* scraping, which is performed according to more or less restrictive

collection criteria (e.g., all URLs within a specific domain, content written in a specific language or covering a specific topic), and *untargeted* scraping (such as web crawling without collection criteria), where the software follows each discovered link on an unrestricted basis. Untargeted scraping can result in the exploration of a large portion of the internet, increasing the risk that the controller has limited knowledge of what personal data are collected.

II. Key points of the Guidance

The Guidelines address the qualification of organizations as controllers or processors, the data processing principles under Art. 5 GDPR (lawfulness, purpose limitation, transparency, data minimization and accuracy), the legal basis for the processing under Art. 6 GDPR, and the treatment of special categories of personal data under Art. 9 GDPR.

1. Applicability of the GDPR

The EDPB recalls that the GDPR applies to web scraping when it includes the processing (e.g., extraction, cleaning, structuring and storing) of personal data, i.e., data that directly or indirectly identifies a natural person considering all the means reasonably likely to be used (Art. 4(1), Recital 26 GDPR). For mixed datasets (including both personal and non-personal data), the GDPR applies to the processing of the personal data contained therein. The EDPB recognizes the challenges for controllers to meet GDPR requirements in the context of data scraping, especially to determine whether and what types of personal data are affected, and related to which data subjects. However, the EDPB stresses that organizations nevertheless bear responsibility to demonstrate compliance under the accountability principles (Art. 5(2) GDPR), and points to the significant data protection risks associated with web scraping, including risks posed by AI models potentially memorizing and reproducing personal data included in training data sets.

2. Controller, joint controller and processor roles

The EDPB states that the organization performing the scraping is not necessarily the controller under the GDPR. The qualification of the organizations involved as controllers, joint controllers or processors should, in the EDPB's view, be analyzed on a case-by-case basis, based on the factual

circumstances of the processing (cf. EDPB guidelines 07/2020 on the concepts of controller and processor in the GDPR).

In the EDPB's view, where an AI developer contracts a scraper to create a training dataset according to documented instructions in particular on data sources and categories, the scraper may be regarded as a processor, and the AI developer as the controller. Where an AI developer trains a model using a dataset already scraped by another entity, each party is, in principle, responsible for its own, separate processing activities, since each determines the objectives and essential means of its own processing. The EDPB considers that the scraper is not, in principle, responsible for the re-use of the data. However, joint controllership may arise where the scraper and the AI developer jointly determine the purposes and means of the processing. This may be the case where two organizations jointly develop an AI model, even if one organization shall carry out the web scraping and preparation of the training data set, and the other organization develops the AI model based on this data set.

3. Lawfulness and purpose limitation

Personal data must be processed lawfully, fairly and in a transparent manner. For assessing the lawfulness of the processing, the EDPB refers to a consideration of the datasets (containing the scraped data) that will be used, and/or the websites that should be scraped. Personal data must also be collected for specified, explicit and legitimate purposes and must not be further processed in a manner incompatible with those purposes (Art. 5(1)(b) GDPR). The EDPB refers, for further considerations on purpose limitation in the context of AI development, to its opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024.

4. Transparency

Personal data must be processed in a transparent manner in relation to the data subject, which includes informing the data subject about the processing of their personal data. The EDPB acknowledges that identifying and informing the data subject effectively is often difficult,

impracticable or, even, objectively impossible when scraping large amounts of data from the internet, which makes compliance with the principle of transparency challenging.

Individual information of data subjects is not required where provision of such information proves impossible or requires disproportionate effort (Art. 14(5)(b) GDPR). The EDPB stresses that the exception should not be routinely relied upon where applied in scenarios other than for purposes of archiving in the public interest, for scientific or historical research purposes or statistical purposes. The EDPB clarifies that a controller seeking to rely on this exception should balance – with respect to the dataset as a whole – its effort to provide the information to the data subject against the impact and effects on the data subject if they were not provided with the information, taking into account the number of data subjects, the age of the data, and any appropriate safeguards adopted.

Notably, the exception is not available by default. The EDPB gives the example of a provider that, to fine-tune an LLM, collects personal data from a closed user group of 5,000 individuals discussing a specific topic over the last two years, who are directly identifiable through their full names, email addresses and unique participant IDs. In such a case, the EDPB considers informing the data subjects individually as not involving disproportionate effort, and therefore requires the controller to do so. Where the dataset is limited, the data is recent and the data subjects identifiable and contactable, individual information remains required.

Where the controller relies on this exception, it must in any case take appropriate measures to protect the data subject's rights, freedoms and legitimate interests. One measure that controllers must always take is to make the information publicly available, for instance by publishing a privacy policy and/or a specific notice on the controller's website. The EDPB points out strict requirements for such information: The information should indicate in particular whether the sources used for data scraping are publicly accessible or non-publicly available and, where the data is crawled and scraped from online sources, include the crawler's characteristics. Where providing a complete list of sources would itself be impossible or require disproportionate effort, the sources should be listed to the greatest extent possible, together with an indication of the types of sources not included and the reasons for not including them. The EDPB considers it good practice to include the domain names and URLs of the scraped web pages, to provide them in a searchable format, and to include the date

or period of collection. Data subjects should further be informed about their rights, and how to exercise them.

The EDPB mentions further measures (in addition to making the information about data processing publicly available) where the controller seeks to not individually inform affected data subjects, including:

- making a data protection impact assessment (Art. 35 GDPR) publicly available,
- anonymizing or pseudonymizing the relevant personal data,
- Minimizing data collection and storage periods, and
- enhancing the level of data security by implementing technical and organizational measures.

Where an organization obtains already scraped datasets to develop AI models from a third party, the EDPB recommends also providing the means to contact the controller from whom the data was obtained. The EDPB considers it a good practice to link directly to the website of the controller that performed the web scraping, accompanied by a clear, concise explanation of the conditions under which the personal data was collected, in clear and plain language.

5. Data minimization

The EDPB considers that the principle of data minimization (Art. 5(1)(c) GDPR) does not prohibit training an AI model with large volumes of data. However, the EDPB requires the controller not to process personal data that would not be necessary, relevant or adequate for the development of the AI model. Therefore, the Guidelines set out specific safeguards to be taken at different stages of the processing:

Before the collection, the controller should:

i. consider using synthetic data instead of personal data;

ii. define precise collection criteria;

iii. undertake a data mapping and inventory process;

iv. apply filters to exclude unnecessary categories of data (e.g., bank transaction data, location/geolocation data);

v. exclude from collection certain categories of websites, such that structurally contain data concerning vulnerable persons (e.g., minors) or data of specific sensitivity (e.g., financial information, location information); and

vi. exclude from collection websites that clearly oppose scraping through technical measures such as robots.txt, ai.txt files, or CAPTCHA.

During and after the collection, the controller should:

vii. apply syntax-based filtering mechanisms (e.g., regular expressions) to filter out personal data identifiable by format (such as social security numbers or telephone numbers);

viii. where feasible, replace real data with synthetic data; and

ix. where feasible, anonymize or pseudonymize personal data.

6. Accuracy

Art. 5(1)(d) GDPR requires personal data to be accurate and, where necessary, kept up to date. The EDPB states that this requirement not only applies to the personal data collected and processed in the context of the development of an AI model, but also to any output of such AI model qualifying as personal data. Therefore, the EDPB requires the controller developing an AI model to also consider any output produced in the context of any training and development of the AI model.

The EDPB recognizes that if data is obtained from different and/or unreliable sources, ensuring accuracy may be difficult or even impossible. Therefore, the EDPB recommends that controllers

scrape from reliable sources (avoiding secondary or outdated aggregators), timestamp data to be able to check whether it is up to date, and validate the data before using it in AI training.

7. Lawful basis

Before carrying out scraping operations involving the collection of personal data or re-use of scraped data, the controller must identify a lawful basis under Art. 6(1) GDPR. The EDPB observes that legitimate interest (Art. 6(1)(f) GDPR) is the legal basis most commonly relied on for web scraping in the context of generative AI training by private entities, while the other grounds under Art. 6(1) GDPR will generally be less likely to apply.

a) Relevant lawful basis

The EDPB considers that consent is unlikely to constitute a valid basis for web scraping, since organizations scraping data from the internet do not have a direct relationship with data subjects and are most probably unable to identify and obtain consent from each individual before scraping. The EDPB further notes that the mere fact that data subjects have made their personal data publicly available online does not mean that they consented to the scraping of that data for a specific purpose. The absence or non-applicability of a robots.txt file on a website does not amount to consent.

b) Legitimate interest (Art. 6(1)(f) GDPR)

In relation to Art. 6(1)(f) GDPR, referring to its guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR and its opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, the EDPB recalls that three cumulative conditions must be fulfilled: (i) a lawful, clearly defined and real legitimate interest; (ii) the necessity of the processing for the purposes of the legitimate interests pursued; and (iii) a the balancing test.

i. **Legitimate interest**: The interest must be lawful, clearly and precisely articulated, and real and present, not speculative. The EDPB has previously identified examples in the AI context, including developing conversational agents, detecting fraudulent content, and improving threat detection.

For general-purpose AI models, the EDPB recommends indicating the objective pursued by the development.

ii. **Necessity**: Controllers must assess whether scraping is suitable and necessary for the purpose, and whether less intrusive alternatives exist. Narrow collection criteria and pseudonymized or synthetic data may be important safeguards.

iii. **Balancing test**: The balancing test requires weighing the data subjects' interests, fundamental rights and freedoms against the legitimate interests of the controller or a third party. The EDPB identifies, in particular, the following elements as relevant:

- Data subjects' interests, fundamental rights and freedoms: Relevant interests include the interest in self-determination and retaining control over personal data published online for a different purpose. Large-scale, indiscriminate collection may create a "chilling effect" on freedom of expression due to self-censoring, and the generation of deep-fakes adds to the impact on data subjects' rights and the right to dignity. In the absence of sufficient safeguards, web scraping may lead to significant impacts due to the large volume and indiscriminate nature of the collection.

- Data processing's impact on data subjects: Web scraping may capture, in addition to special categories of personal data and criminal data, data revealing highly private information, such as location or financial data, as well as personal data about minors. The more sensitive or private the nature of the data, the more likely the processing will have a negative impact on the data subject, and the more weight it should be attributed in the balancing test. Furthermore, the impact on the data subject is influenced by the particularities of web scraping, such as processing a large number of data subjects and volume of data, which may have been published long time ago. Looking at the consequences of the processing, it is difficult for data subjects to oppose the scraping of their data or to make them unreachable to scrapers. Moreover, the current technical state of the art does not allow that personal data are easily deleted from a generative AI model once the model is trained.

- Reasonable expectations: Data subjects may be aware that data they publish online may be

reused by third parties but cannot always be expected to anticipate such processing for all purposes and all types of data. The mere fulfilment of the information obligations under Arts. 12–14 GDPR is not sufficient to establish reasonable expectations. Relevant factors include the nature of the source website, whether the data subject made the data publicly accessible (where the data subjects are more likely to reasonably expect their data being processed by others), any anti-scraping measures on the website (e.g., robot.txt, ai.txt or CAPTCHAs), and the characteristics of the data subjects.

- Mitigating measures: Where data subjects' interests override the controller's legitimate interests, the controller may introduce mitigating measures to enable a processing based on Art. 6(1)(f) GDPR. The EDPB provides a non-exhaustive list of potential technical or organisational measures, such as excluding websites accessible only when logged in or certain data categories or publications potentially including personal data entailing risks for particular (groups of) data subjects, limiting collection based on time periods, publishing updated lists of scraped websites, implementing a prior (unconditional) objection mechanism and/or opt-out list from certain websites or online platforms, deleting or anonymizing unnecessary personal data, and implementing measures against memorization and regurgitation.

The EDPB illustrates the balancing test with two examples: collecting publicly available voice recordings to develop a voice generation tool without any additional measures cannot, in the EDPB's view, be based on legitimate interests. Conversely, a text-based system using only freely accessible sources, excluding copyrighted content (i.e., using only content in the public domain or for which rights holders have not objected to text and data mining as permitted under Directive 2019/790), implementing memorization safeguards, facilitating the exercise of data subjects' rights, and publishing a comprehensive privacy policy may meet the balancing test.

8. Special categories of personal data

The EDPB reinforces that, in general, any processing of special categories of personal data which cannot be based on a derogation pursuant to Art. 9(2) GDPR in addition to a legal basis under Art. 6(1) GDPR is unlawful and cannot be justified under EU data protection law.

However, where there is no intended scraping of special categories of personal data, the EDPB considers that for any incidental and residual collection of such types of data, the CJEU's ruling in *GC and Others* (CJEU, judgment of 24 September 2019, C-136/17) in the context of the activity of a search engine may potentially be applied. In its ruling, the CJEU held that although there is no general exemption from the prohibitions in Art. 9 or Art. 10 GDPR, the specific features of the processing carried out the operator of a search engine may have an effect on the extent of the operator's responsibility and obligations; the prohibition in Art. 9(1) GDPR applies "*within the framework of his responsibilities, powers and capabilities*" (para. 45 seq.), and therefore only by reason of referencing and via verification on the basis of a request by the data subject.

However, the EDPB emphasizes that this exemption cannot be broadly applied on a general basis. Rather, the EDPB requires the controller to carry out a case-by-case analysis to assess whether:

i. the processing activity has relevant similarities with the processing activity of a search engine as referred to in the judgment;

ii. the processing includes only the incidental and residual processing of special categories of personal data, and no intentional processing;

iii. it is difficult or impossible to assess whether and to what extent the processing includes special categories of personal data; and

iv. the controller implements measures "within the framework of his responsibilities, powers and capabilities" to prevent the collection and the dissemination of special categories of personal data.

Regarding condition (iv.), the EDPB specifies the minimum measures that the controller's "*responsibilities, powers and capabilities*" will include:

- before the collection of the data, defining precise criteria and applying filters to prevent collection of special categories of personal data, and excluding categories of websites that structurally contain such data;

- after the collection of the data, deleting any special categories of personal data from datasets immediately or as soon as it is identified, including upon request from a data subject where the request constitutes a plausible indication that the data falls under Art. 9(1) GDPR;

- during the development of the AI model, preventing extraction of special categories of personal data from the model, providing assurance regarding state-of-the-art resistance to privacy attacks, and applying output filters; and

- after the development of the AI model, implementing a process to constantly monitor the output and immediately taking measures where special categories of personal data are nevertheless produced. Model unlearning or other processes with equal effect should be considered as they become available through technological progress.

As part of the accountability principle, the controller should be able to demonstrate that conditions (i.) – (iii.) are met and that the measures under condition (iv.) are relevant and effective. The measures should be verified and assessed on a regular basis, and must be in line with state-of-the-art techniques which should be constantly monitored.

III. What's next?

The Guidelines have been adopted for public consultation until 30 October 2026. Companies engaged in web scraping or using data sets scraped from a third-party provider for generative AI training should review the Guidelines alongside the EDPB's existing guidance, in particular EDPB guidelines 1/2024 on legitimate interest and EDPB Opinion 28/2024 on AI models. Although the

Guidelines are non-binding and remain subject to potential change following the consultation period, they provide important indications of how European data protection authorities are likely to assess web scraping practices.

Companies should in any case consider the EDPB's positions when reviewing their current scraping practices, in particular as regards the EDPB's expectations on the exclusion of certain data categories and sources, the publication of comprehensive and searchable source lists, the important role of mitigating measures under the legitimate interest balancing test, and the phase-specific measures the EDPB considers necessary for preventing the collection and dissemination of special categories of personal data.

In addition, companies engaged in web scraping activities and the subsequent development of AI models must not lose sight of further challenges resulting from various other areas of law, including not least copyright law requirements and the related developments to text and data mining.

Authored by Martin Pflueger, Henrik Hanssen, Sebastian Eisenberger, Alexander Bathelt, and Charlotte Haddad.

Contacts

Dr. Martin Pflueger

Partner

Munich

Email me

Dr. Henrik Hanssen

Counsel

Hamburg

Email me

Sebastian Eisenberger

Senior Associate

Munich

Email me

Dr. Alexander Bathelt

Associate

Munich

Email me

Charlotte Haddad

Associate

Paris

Email me