# Web scraping for generative AI: How and when to use it

EDPB European Data Protection Board

GUIDELINES AT A GLANCE

October 2026

Web scraping is a commonly used technique for extracting large amounts of data from publicly available web services. This technique can be a powerful tool for generative AI development, but it poses significant risks as it typically operates 'invisibly'.

Generative AI is a technology designed to create new content by learning patterns from existing data. It employs specialised machine learning models capable of producing diverse outputs, including text, images, and audio.

The **EDPB guidelines on web scraping** in the context of generative AI clarify how to comply with the GDPR, covering both legal and technical aspects.



## Common web scraping scenarios for AI training

In the context of AI training, an AI developer may employ web scraping not only to train new generative AI models but also to fine-tune existing models for specific tasks.

The guidelines cover these common situations:

- Scraping data from the internet themselves or by contracting another party.
- Obtaining and re-using a dataset that has already been scraped by another organisation (data broker).

# Allocation of responsibilities in AI data scraping

The GDPR applies to web scraping when it includes personal data processing operations, such as **collection, storage, organisation** and **retrieval**. Organisations may be involved in the data collection process for the creation of an AI training dataset, with varying degrees of involvement:

- **When you (AI developer) hire another party to scrape data for you**, you typically decide how and why scraping happens, making you the ***data controller*** under the GDPR. If the scraper creates a training dataset on your behalf, according to your documented instructions, they may be regarded as a ***processor***.
- **When you make use of pre-scraped data**: The scraper and you are, in principle, responsible for your own, separate, processing activities, since each of you determines the objectives and the essential means of your own processing. The scraper is not, in principle, responsible for the re-use of the data.
- **When the scraper and you determine the purposes and means of the processing together**, you are jointly responsible for the processing.

# Practical example

Here is a specific scenario extracted from the Guidelines:

**Example**

**Controller/joint controllers/processor (section 3.2.1, paragraph 20, page 9, example 1)**

**Context:** Two organisations make a joint decision to develop an AI model. Both parties agree that one of the organisations will scrape data from the web and create data sets for the development of the AI model, while the other organisation develops the model, using the data sets created by the scraper.

**Assessment:** The collection criteria for the scraping of personal data are determined jointly by both parties. Even though they perform separate processing activities, they determine the purposes and means of the processing (the web scraping and the development of the model) together and therefore are jointly responsible for the processing of personal data.

# Processing personal data: Legal basis and key principles

Organisations often use "legitimate interest" (Art. 6(1)(f) GDPR) as the legal basis for scraping data to train AI models. You can find more detailed information about when and how to apply this legal basis **here**.

Particular attention needs to be paid to the following data protection principles:

## Lawfulness

You must ensure personal data is processed **legally, fairly** and **transparently**. This means carefully checking any datasets you use or websites you scrape, and making sure your scraping purpose is **lawful**.

## Purpose limitation

Personal data should be collected for **specified, explicit** and **legitimate purposes** and used only for those purposes.

This can include the processing of personal data to avoid the risks of potential biases and errors when this is clearly and specifically identified within the purpose, and the personal data is necessary for that purpose (e.g. they cannot be effectively achieved by processing other data, including synthetic or anonymised data).

## Transparency

You should inform individuals about how their personal data is processed, what rights they have under the GDPR, and how they can exercise those rights. Information should be easy for users to find and understand.



However, depending on how the data processing is precisely designed, **you might not have to inform users individually** about the web scraping **if this proves to be impossible or require excessive effort**. For example, this may apply when you need to collect a very large dataset from public social media pages or forums, including very old content, and you cannot directly contact the individuals concerned. In such cases, you should still provide the information in another effective way, for example by publishing it on your website in a clear, plain and easy-to-find notice. You should also take appropriate protective measures, such as excluding data from web pages where individuals are directly identifiable.

# Practical example

Here is a specific scenario extracted from the Guidelines:

**Example**

**(section 3.2.2, paragraph 34, page 11-12, example 3)**

**Context:** An AI provider intends to further train a language model that generates text from short prompts. For this, it collects personal data from a closed group of 5,000 individuals who have discussed a specific topic over the last two years using a particular business or social jargon. Individuals can be directly identified by their full names, email addresses and unique participant IDs.

**What to do:** Since the data is recent and the provider can contact the individuals concerned, providing the information required under Art. 14 GDPR would not involve disproportionate effort. The provider should therefore inform each person individually about the scraping.

## Data minimisation

You must **collect only the personal data necessary** for the purpose of the processing.

Before scraping begins, you should **assess what data is truly needed** for the development of the model to ensure you do not collect unnecessary personal data. More specifically:

| |
|---|
| **Consider using synthetic data instead of personal data.** |
| **Define precise collection criteria.** |
| **Undertake a data mapping and inventory process.** |

**Apply filters** to exclude the collection of certain categories where they are not necessary (e.g. bank transaction data, location/geolocation data, etc.).

**Exclude from the collection** certain categories of websites (e.g. sites or social networks used mainly by minors), including websites which contain sensitive personal data like financial or location information.

**Avoid websites that block scraping** with technical measures (such as robots authentication to view content, robots.txt or ai.txt files, or CAPTCHA) aimed at ensuring that an action can only be carried out by a human being.

### Accuracy

Personal data processed **must be accurate and up to date for its intended purpose.** You should scrape data only from **reliable sources**: when possible, choose sources that are official, trustworthy, and maintained, and avoid secondary and outdated aggregators.

The date and time of collection should also be recorded so it is clear how old the current the data is.

In addition, check the data before using it for AI training, for example by making sure email addresses and phone numbers are in the correct format and by reviewing random samples to confirm the information is accurate.

## Special categories of personal data

Processing special categories of personal data, such as data revealing racial or ethnic origin or political opinion, and concerning health or sexual orientation, is in principle **prohibited.** If web scraping involves such data, both a lawful basis under **Art. 6 GDPR and an exception under Art. 9(2) GDPR are required.**

While organisations should prevent collecting sensitive personal data, it is not always possible to know for sure whether collection involves this kind of data until after scraping has occurred.

Even with proper safeguards, organisations may accidentally collect sensitive personal data during AI training. Organisations should assess these cases individually and be able to show that 1) the processing is similar to that of a search engine, 2) the sensitive data is collected only incidentally and not intentionally, 3) it is difficult or impossible to know in advance whether such data will be collected, and 4) appropriate safeguards are in place to prevent its collection and dissemination.

This document provides a simplified overview of the guidelines. For more comprehensive legal explanations and examples, please consult the full text of the guidelines.

Read the complete guidelines