BCLP. Client Intelligent

Insights

WEB SCRAPING FOR AI TRAINING IN FRANCE

MANAGING DATA PROTECTION AND COPYRIGHT COMPLIANCE

Nov 03, 2025

SUMMARY

In the age of online information and the rise of artificial intelligence, web scraping has become a widespread method for feeding and training AI systems. However, this proliferation presents major legal risks, particularly concerning data protection and intellectual property, including potential privacy violations and infringement of copyright and neighbouring rights.

The existing French regulatory framework lacks specific rules directly applicable to web scraping. We anticipate that more express, sector-specific regulation will emerge to govern web scraping practices, particularly for AI model training purposes, striking a balance between technological innovation and respect for the fundamental rights of content creators whose materials are ingested by AI models.

With the EU AI Act's transparency requirements for generative AI models which mandate detailed summaries of data sources used for model training and EU copyright law compliance policies, we expect increased tensions between rights-holders and AI model developers.

For those operating in France, useful guidance can be gleaned from regulatory recommendations[1].

RECONCILING WEB SCRAPING WITH PERSONAL DATA PROTECTION

UNDERSTANDING THE RISKS

Web scraping poses significant risks to individuals, including mass collection and excessive accumulation of large amounts of data, invasion of privacy, and processing of sensitive data.

REGULATORY ENFORCEMENT

The French data protection authority (CNIL) has already imposed severe penalties on companies using data available online. In its decision against Clearview[1], the American company had collected large volumes of photos available online via web scraping to feed its facial recognition system. The CNIL ruled that the collection and processing of biometric data without consent or other valid legal basis was intrusive, massive, and disproportionate to the commercial interests involved.

More recently, the CNIL sanctioned the company Kaspr[2], which had built a database of information collected from the LinkedIn social network. The CNIL considered that the processing lacked a valid legal basis under the GDPR since the company collected contact details of people who had restricted the visibility of their data on LinkedIn, thereby exceeding users' legitimate expectations.

LEGITIMATE INTEREST AS A LEGAL BASIS

Some national data protection authorities argue that web scraping for AI model training should only be permitted with the prior consent of the individuals concerned. However, the CNIL does not share this position and refuses to impose such a generalised requirement, particularly given the technical difficulty of obtaining valid consent in the context of scraping, which would make access to training data far too complex, if not impossible.

It is important to note that consent cannot be inferred from the mere fact that a person has made their personal data public. Furthermore, the fact that sensitive data is accessible online does not necessarily mean that it has been made public by the data subject in order to invoke the exception to the prohibition on processing such data provided for in Article 9.2(e) of the GDPR.

According to the CNIL, the following interests could be considered prima facie legitimate in the course of developing an AI system:

- conducting scientific research;
- facilitating public access to certain information;
- developing new systems and features for users of a service;
- improving a product or service to increase its performance;
- and developing an AI system to detect fraudulent content or behaviour.

The extent of the expected benefits for the data controller, but also for third parties such as the end users of the AI system or the public more generally, could enable a web scraper to identify a sufficient legitimate interest to justify the activity.

The CNIL points out that one of the conditions necessary to characterise the legitimate interest of the data controller in web scraping is the need to demonstrate that the data subject can reasonably expect that the data they publish will be processed at a later date. When assessing whether data subjects expected their data to be reused, the data controller may consider:

- the publicly accessible nature of the data;
- the nature of the source websites (social networks, online forums, etc.);
- the absence of restrictions imposed by the websites being scraped (for example, in the terms of use, or through technical measures such as exclusion protocols like robots.txt, or blocking measures such as CAPTCHAs);
- the type of publication (for example, an article published on a freely accessible blog is not private, whereas a post on a social network published with access restrictions may be considered private by internet users);
- and the nature of the relationship between the data subject and the data controller.

DATA MINIMISATION PRINCIPLE

When an AI tool performs high-volume and indiscriminate scraping on the web, it is unlikely that all the personal data collected will be relevant to the processing carried out.

The CNIL therefore requires the data controller to define precise collection criteria in advance using filters. When this is not possible, it recommends that certain types of sites (e.g., social networks used mainly by minors or sites containing sensitive data) are simply excluded from the collection exercise.

If the collection process has incorporated irrelevant data, it is necessary to delete it and apply anonymisation or pseudonymisation processes to the retained data. It is also advisable to prevent any cross-referencing of data based on individuals' identifiers, for example by replacing them with random pseudonyms specific to each piece of content.

INFORMATION AND EXERCISE OF DATA SUBJECT RIGHTS

Providing individual information to data subjects is often complex or even impossible when collecting web scraped data, since finding ways to contact individuals may require collecting additional or more identifying data.

In this case, the CNIL suggests publishing a comprehensive general information notice on the website of the data controller.

The CNIL suggests disseminating information on data collection and individuals' rights as widely as possible by using multiple media (e.g., online articles, the data controller's social media accounts), publishing an updated list of sites affected by web scraping, and even disseminating this information through the publisher of the site where the data was collected.

The CNIL also suggests providing for a discretionary and prior right to object by implementing a simple checkbox that is quickly accessible.

In the case of online data collection, the CNIL encourages the development of technical solutions to facilitate compliance with the right to object prior to data collection. In addition to opt-out mechanisms (identical to those implemented in the field of intellectual property, as discussed in the second part of this article), "rejection lists" containing the identities of individuals who refuse to allow their personal data to be used could be implemented.

Respect for individuals' rights requires data controllers to adopt a somewhat paradoxical approach, as they must simultaneously limit the personal data collected whilst putting in place measures to facilitate the identification of individuals so that they can exercise their rights in relation to their data.

For example, in the case of a set of image data created from web scraping of freely accessible data online on a limited number of websites, it is suggested that the display name and URL of the source of each image be retained to facilitate the identification of individuals.

It should be noted that the exercise of rights over an AI model is not absolute. The proportionality of a data deletion demand depends on the sensitivity of the data and the risks that its regurgitation or disclosure would pose to individuals, balanced against the infringement of the organisation's freedom to conduct business.

Even when web scraping complies with the GDPR, it may still be illegal if it infringes on the rights of the authors of the collected content.

RECONCILING WEB SCRAPING WITH COPYRIGHT

Training generative artificial intelligence systems requires the processing of content that may be subject to copyright.

To enable the training of these systems and the capture of protected content, the provisions introduced by Articles 3 and 4 of Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market have recently been activated.

ARTICLE 3 EXCEPTION

The exception to the authors' monopoly stated under Article 3 applies only to "*research organisations and cultural heritage institutions*" acting for scientific research purposes.

Users must have "*lawful access*" to the works, and rightsholders cannot opt out. It authorises, without compensation, the automated reproduction of works for scientific research purposes, but only by public or non-profit organisations.

ARTICLE 4 EXCEPTION

The exception under Article 4 is open to any user, including commercial entities. Rightsholders may "*expressly reserve their rights in an appropriate manner*," for example through machine-readable metadata, contract terms, or technical measures.

On 10 July 2025, the final version of the Code of Best Practices was published, which provides guidance as to what measures constitute an "*opt out*" expressed by machine-readable means: crawlers compatible with the robots.txt protocol; and other appropriate exclusion protocols (e.g., metadata by asset or location), which have been adopted by standardisation bodies or widely implemented by the cultural sectors concerned.

As noted above, the EU AI Act requires providers of general-purpose AI (GPAI) models to put in place a policy to comply with EU law on copyright and related rights and implement a methodology to ensure rights reservations under Directive 2019/790 are appropriately respected when a GPAI model is trained and deployed.

PENDING LITIGATION

The scope of Article 4 of Directive (EU) 2019/790 is currently being examined in Like Company v. Google Ireland (C-250/25), the first generative AI copyright dispute referred to the Court of Justice of the European Union. The ruling is expected to clarify how EU copyright law applies to generative AI training practices.

AI system developers may therefore use web scraping or data mining to train their models, but they must comply with strict limitations, implement technical measures to ensure compliance with data protection and copyright regulations, and document them.

---

[1] CNIL decision SAN-2022-019 - October 17, 2022

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

[2] CNIL Decision SAN-2024-020 - December 5, 2024

RELATED CAPABILITIES

- Data Privacy & Security

- Data Privacy, Telecommunications & Collections
- General Data Protection Regulation

MEET THE TEAM

Pierre-Emmanuel Froge

Paris

pierreemmanuel.froge@bclplaw.com
+33 (0) 1 44 17 76 21
