Hogan Lovells

News

Development of an AI system: CNIL issues guidelines regarding collection of data via web scraping

27 June 2025

On 19 June 2025, CNIL published two additional “how-to-sheets” on artificial intelligence, one on the legitimate interest and the other on the collection of data via web scraping. These documents aim to clarify the rules applicable to the creation of training datasets containing personal data.

On June 19, 2025, CNIL published two additional “how-to-sheets” on artificial intelligence. The first one sets out the conditions under which the legal basis of legitimate interest may be used for the development of an AI system (see here our post), while the second focuses specifically on the collection of data via web scraping.

In its second "how-to-sheet,” CNIL details the measures to be taken for proper collection of data via web scraping. The widespread use of web scraping has indeed significantly changed how the Internet is used, making all data published online potentially accessible, collectable, and reusable. This

practice raises significant risks for data subjects, including:

- **Infringements of privacy and other rights protected under the GDPR** (due to the volume of data collected, the sensitivity of certain data, the inclusion of vulnerable individuals, and challenges in exercising rights);
- **The risk of unlawful data collection**, particularly where intellectual property rights or consent requirements are at stake; and
- **Impacts on freedom of expression**, including chilling effects or self-censorship.

Although web scraping is not prohibited per se, the CNIL emphasizes the need for a case-by-case assessment and calls for the implementation of appropriate safeguards. It also recommends the introduction of specific legislation to regulate scraping practices by public authorities. In the absence of such a framework, the CNIL reminds controllers of their obligations and sets out the conditions under which these practices may be used for training AI systems.

Complying with mandatory measures under the GDPR

The CNIL reiterates that certain measures are mandatory, particularly under the data minimization principle (Article 5.1(c) of the GDPR). This includes:

- **Defining specific collection criteria** in advance;
- **Excluding unnecessary categories of data**, such as banking or geolocation data, using filters;
- Where filtering is not feasible, **excluding certain types of websites**, such as those predominantly used by minors or that structurally contain sensitive data or information on vulnerable individuals;
- **Deleting non-relevant data collected in error** as soon as they are identified;

- **Excluding websites that clearly object to scraping**, through mechanisms such as robots.txt files or CAPTCHA barriers.

The CNIL emphasizes that special attention must be paid to sensitive data, given the large volumes typically involved. Residual and unintentional collection of such data, despite precautions, is not unlawful per se, as confirmed by the CJEU (Case C-136/17). However, once a controller becomes aware that it is processing sensitive data, it must ensure immediate deletion, using automated means where possible.

Additionally, the CNIL recalls that processing sensitive data is only permitted by way of exception, notably where the data have been manifestly made public by the data subject. This requires a clear, positive action by the individual, made knowingly (CJEU, Case C-252/21, Meta Platforms).

Respecting reasonable expectations

To ensure the balance required under the legitimate interest basis, the controller must also take into account the reasonable expectations of data subjects. In this respect, the CNIL refers to the following criteria:

- **Nature of the relationship** between the data subjects and the controller;
- **Explicit restrictions imposed by websites** (e.g. T&Cs, robots.txt, CAPTCHA): failure to comply with such restrictions means that the processing does not meet reasonable expectations;
- **Nature of the source website** (e.g. social media, forums);
- **Type of content** (e.g. public blog post vs. restricted social media post);
- **Public accessibility** of the data (or lack thereof).

Implementing additional safeguards

Finally, the CNIL stresses that the controller will generally need to implement additional safeguards to mitigate the impact on data subjects' rights and freedoms, particularly in view of the intended use of the trained AI system and its actual impact on data subjects. The controller must assess on a case-by-case basis whether such safeguards are required, depending on the specific modalities of the processing. Recommended safeguards include:

- **Establishing an exclusion list of websites by default**, especially those hosting sensitive or highly intrusive data (e.g. health forums, pornographic websites, genealogical databases);
- **Excluding websites that oppose scraping**, either through their terms of use or through explicit technical measures;
- **Limiting collection to data that is freely accessible** and that individuals are aware they are making publicly available;
- **Providing broad information to data subjects**, using various channels (e.g. website, social media, published list of scraped websites, or in collaboration with website publishers);
- **Providing a prior and effective right to object**, supported by appropriate technical mechanisms (e.g. opt-out systems, suppression lists);
- **Applying anonymization** or, failing that, **pseudonymization measures** immediately after collection;
- **Preventing data reidentification or linkage using identifiers**, unless such linkage is demonstrably necessary, for example.

Authored by Joséphine Beaufour and Julie Schwartz.

Contacts

Joséphine Beaufour

Senior Associate

Paris

Email me

Julie Schwartz

Counsel

Paris

Email me