30 October 2026

Publicly available data and web scraping

Executive summary

Web scraping is a method currently used for different use-cases, from market research for competitive pricing, to supply chain and logistics analysis, sentiment analysis, lead generation or to the development of AI models and systems. Just as for the upcoming AI Act and General Data Protection Regulation[1] guidance, we recommend that the European Data Protection Board (EDPB) work in close collaboration with the AI Office. The Guidelines[2] should not unnecessarily prevent different business use cases. Instead, the Guidelines should measure the necessity and proportionality of data collection based on the purpose, scale, sources and intended use of the data. We recommend that the Guidelines on web scraping in the context of generative AI prioritise the following:

- Better guide companies in balancing the amount of information which should be shared with data subjects under Art. 13 GDPR for web-scraping and recognise cases where Data Protection Impact Assessments (DPIAs) cannot be shared;
- Proportionally apply the principles under Art. 5 GDPR, in particular data minimisation, without setting new additional obligations;
- Recognise cases where legitimate interest applies, including where the technology may evolve over time.

[1] Regulation (EU) 2016/679.

[2] Draft Guidelines 03/2026 on web scraping in the context of generative AI.

DIGITALEUROPE
Rue de la Science, 37, B-1040 Brussels
+32 2 609 53 10 ▶ Info@digitaleurope.org
▶ www.digitaleurope.org
EU Transparency Register: 64270747023-20

Table of contents

Executive summary **1**
Table of contents **2**
Coherence with the Digital Omnibus **3**
Information obligations **4**
Proportionate and technologically realistic data minimisation ... Error! Bookmark not defined.
Legitimate interest ..... Error! Bookmark not defined.

Coherence with the Digital Omnibus

The Web Scraping draft Guidelines were proposed in the midst of the simplifications efforts under the Digital Omnibus. Under the omnibus, the proposed amendments to Art. 9 could help clarify that the processing of special categories of personal data may be lawful where necessary and subject to appropriate safeguards. The new Article 88c, would also help clarify that the development and operation of AI systems and models may rely on legitimate interest.[3]The Guidelines should also align with Art. 10 of the AI Act on data governance, and with Art. 53 of the AI Act, on the distinct obligations of general-purpose AI model providers and deployers.

Information obligations

The Digital Omnibus helpfully proposes to harmonise Data Protection Impact Assessments (DPIAs) and solve the issue of differing interpretations and expectations of what a DPIA should include and its format. We support common **voluntary** DPIA templates, developed by the EDPB in coordination with stakeholders.

One of the biggest challenges with the information obligation under the GDPR is to balance the need for information to be concise, to be understandable by the data subject, but also detailed enough to be complete. This difficulty is expanded in the context of web-scraping, in particular where data subjects need to be informed individually.

The obligation to provide and disclose characteristics of web crawlers and a list of sources, including domain names URLs, and collections dates, would in some cases lead to data subjects receiving a very long list of technical information, but it could also entail risks of re-identification or data poisoning and lead to cybersecurity risks for individuals.[4]

Instead of requiring transparency of all sources web scraped by an organisation, we suggest that under some circumstances in highly critical sectors, such as cybersecurity, sufficient transparency could be demonstrated through publication of source categories, collection methodology and representative examples. This aligns with a similar approach taken within Art 14(1)(e), where controllers must provide the categories of personal data concerned rather than listing everyone.

This is also applicable to sector-specific requirements, as the Guidelines should recognise targeted source selection, provenance, regular updating and filtering as relevant safeguards. The necessity and proportionality of these measures should be assessed in light of the purpose, sources, scale and foreseeable impact.

The draft recommends that Data Protection Impact Assessments (DPIAs) be made publicly available. The draft Guidelines should also recognise that this would go beyond Art. 35 GDPR, and is not always safe, notably as a publication could lead to exposure of commercially sensitive information or security vulnerabilities. Disclosing certain information in the DPIAs risks facilitating attacks on AI models or on the data pipeline.

[3] See: Digital-omnibus-a-first-step-and-what-must-come-next-now.pdf

[4] Para. 31 of the draft Guidelines.

Proportionate and technologically realistic data minimisation

With respect to the principle of data minimisation,[5] the draft Guidelines note that the personal data collected must be adequate, relevant and limited to the purposes of the process-and state that this does not prohibit training an AI model or AI system on large volumes of data. This aligns with the fact that the GDPR does not tie particular obligations to each individual principle. The only specification the GDPR gives is under Recital 39: 'This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum.'

However, the draft Guidelines go a step further and advise that in order to apply the principle of data minimisation, the controller reviews at least six different considerations before personal data is collected through web scraping, and multiple other steps during and after the collection. We recommend the Guidelines clarify these are recommendations come on top of the actual legal obligation under Art. 5 GDPR, notably for data mapping and inventory process, which is not detailed in the Guidelines nor in the GDPR.

The draft Guidelines also suggest that, where feasible, all real data should be replaced with synthetic data.[6] While controllers are required to assess at the time of processing whether the use of training data complies with the GDPR, we caution against mandating replacement or removal of personal data after that stage. Indeed, once training has taken place, often the personal data that is used cannot technically be selectively removed or replaced from a model without full or substantial re-training. Currently, the cost of replacing data with synthetic data would be disproportionate to carry out, as compared to its effectiveness. Companies also need to comply with the data protection by design principle stemming from Art. 25 GDPR.[7]

Further, with regards to notifying data subjects, the Guidelines should distinguish between user-generated content posted on social networking platforms and information intentionally published through government registries, licensing systems, corporate filings, professional credentialing systems and similar transparency mechanisms. Third parties whose personal data appears incidentally in content published by others, such as bystanders in a picture, may have no relationship to the controller, whom would not have contact details or means to identify them. Hence, the Guidelines should avoid treating all scraping operations as representing the same level of risks.

Legitimate interest

The Guidelines refer to the technological developments that have occurred in generative AI over the past years, highlighting that data subjects may be aware that the data they publish could be accessed, collected, and reused by third parties. It is also noted that individuals cannot always be considered to be aware that

[5] Art. 5 GDPR.

[6] Para. 38 of the draft Guidelines.

[7] Guidelines 03/2026 on web scraping in the context of generative AI.

web scraping would take place in all situations. Moreover, the impact of generative AI models, depending on the capabilities of the model, may also change over time.[8]

The draft Guidelines should recognise that legitimate interest under Art. 6 GDPR can be a lawful basis for foundation-model development where the training purpose is sufficiently specified and safeguards materially reduce data-subject risks. They should also acknowledge that legitimate interest may evolve over time with technological developments, but does not require a re-assessment at each step. [9] Recital 47 GDPR already addresses the reasonable expectations of the data subject, and the final Guidelines should further recognise the inherent limitations in predicting all potential consequences. A best-effort approach, offering flexibility within the assessment, would be more practical and aligned with real-world processing scenarios.

The Guidelines should recognise cases where legitimate interest applies, for instance to prevent fraud, strengthen identity assurance, and to detect cybersecurity threats. Reasonable expectations should be assessed in context, including the nature of the source, whether the information was intentionally made public for verification or transparency, the relationship with the data subject, and whether the processing produces a final decision or instead supports authentication or investigation.

FOR MORE INFORMATION, PLEASE CONTACT:

Béatrice Ericson

Senior Manager for AI & Data Policy

beatrice.ericson@digitaleurope.org / +32 492 25 84 32

[8] Ibid.

[9] See: https://www.digitaleurope.org/resources/legitimate-interest-one-of-six-legal-bases-to-process-personal-data/.

THE VIEWS EXPRESSED IN THIS STUDY ARE THOSE OF DIGITALEUROPE AND DO NOT NECESSARILY REFLECT THE POSITIONS OF ITS INDIVIDUAL MEMBER ORGANISATIONS OR CONTRIBUTORS.

About DIGITALEUROPE

DIGITALEUROPE is the leading trade association representing digitally transforming industries in Europe. We stand for a regulatory and investment environment that enables European digitalising businesses across multiple sectors and citizens to prosper from digital technologies. We wish for Europe to grow, attract and sustain the world's best digital talents, investment and technology companies. Together with our members, we shape industry positions on all relevant policy matters, and contribute to their development and implementation. Our membership represents over 56,000 businesses who operate and invest in Europe. It includes corporations and scaleups who are global leaders in their field of activity, as well as national trade associations from across more than 30 European countries.

DIGITALEUROPE
Rue de la Science, 37, B-1040 Brussels
+32 2 609 53 10 ▶ Info@digitaleurope.org
▶ www.digitaleurope.org
EU Transparency Register: 64270747023-20

@DIGITALEUROPE DIGITALEUROPE digitaleurope_org