

Consultation Response: Guidelines 03/2026 on Web Scraping in the Context of Generative AI

Respondent: Jiri Svorc, Solicitor and Independent Legal Consultant

General observations

I welcome the publication of these Guidelines and the EDPB's efforts to provide practical guidance on the use of web-scraped data in the context of generative AI.

In recent years, organisations of all sizes have been trying to navigate an increasingly complex landscape. On the one hand, there is a clear need to protect individuals' rights and ensure compliance with the GDPR. On the other hand, organisations developing or deploying AI systems are facing significant uncertainty regarding what responsible and lawful use of publicly available information looks like in practice.

The Guidelines are therefore both timely and important.

From my own work advising technology companies, SMEs, research organisations and Horizon Europe projects, I have found that the main difficulty is often not understanding the legal principles. Rather, it is applying those principles consistently in situations where the data is publicly available, originates from a wide variety of sources and is used in ways that were unimaginable when much of the information was first published online.

The points below are offered in that spirit.

Comment 1: More guidance on reasonable expectations and different categories of public information

The Guidelines correctly identify the reasonable expectations of data subjects as an important element of the legitimate interests assessment. In practice, however, this remains one of the most difficult concepts to apply consistently. When advising clients, this is often where discussions become most challenging.

One reason for this difficulty is that publicly available information is frequently treated as though it were a single category. In reality, individuals are likely to have very different expectations depending on where information appears online, why it was published and the context in which it is made available.

JIRI SVORC

Solicitor

EMPOWERING INNOVATION
ENSURING INTEGRITY

+420 737160737
+44 (0)7588542026
jiri@jiri.legal
www.jiri.legal

167-169
Great Portland Street
London, W1W 5PF
United Kingdom

Vrchlabska
1131/22
Prague, 19700
Czech Republic

For example, expectations may differ significantly where information appears in an academic publication, a professional biography, a company website, a conference presentation, a public register, an open-source software repository, a discussion forum, or a personal social-media profile.

Information published to support professional activities, academic dissemination or public accountability may raise very different considerations from information shared in a personal or social context. Organisations therefore often struggle to determine how much weight should be given to the fact that information is publicly accessible when carrying out the legitimate interests balancing exercise.

From a practical perspective, organisations would greatly benefit from a more structured framework for assessing reasonable expectations. I would encourage the EDPB to provide additional examples and indicative factors that organisations can consider when carrying out this assessment.

Relevant factors might include:

- the nature of the source;
- the original purpose for which the information was made public;
- whether the context is professional or personal;
- the accessibility of the information;
- the intended audience of the publication; and
- what an average individual would realistically expect to happen to their information.

Greater clarity in this area would significantly improve predictability and consistency in the application of the legitimate interests balancing exercise while recognising the important differences between various categories of publicly accessible information.

Comment 2: More clarity for organisations using third-party datasets and models

Many organisations involved in AI do not scrape data themselves. Instead, they acquire datasets from third parties, license foundation models, fine-tune existing models, or purchase AI-enabled solutions from suppliers. In practice, these organisations are often trying to understand what level of due diligence is expected of them when they have not been directly involved in the original collection of data.

This is an issue I encounter regularly, particularly among SMEs and research organisations that wish to use AI responsibly but have limited visibility of how datasets

were originally assembled. Additional guidance on this point would be extremely helpful.

In particular, organisations would benefit from understanding:

- what questions should reasonably be asked of suppliers;
- what evidence should be obtained;
- what factors increase or decrease the level of due diligence expected; and
- how expectations may differ depending on the level of risk involved.

A practical, risk-based framework would promote accountability while recognising the realities of increasingly complex AI supply chains.

Comment 3: Practical guidance on Article 14 is needed

Transparency remains one of the most challenging issues in this area. In many AI development scenarios, organisations may be dealing with extremely large volumes of information collected from numerous sources over long periods of time. In some cases, identifying and contacting individual data subjects may be impossible in practice. While the Guidelines acknowledge these challenges, additional practical examples would be extremely useful.

In particular, organisations would benefit from more guidance on:

- when Article 14(5)(b) may apply in practice;
- how disproportionate effort should be assessed;
- what evidence organisations should retain to support their assessment; and
- what alternative transparency measures may be considered effective.

Many organisations genuinely wish to be transparent but are uncertain about what that should look like in the context of large-scale AI training activities. Worked examples would therefore be especially valuable.

Comment 4: Encouraging practical technical solutions

Looking ahead, I believe there is considerable value in encouraging the development of practical and interoperable technical mechanisms through which website operators can communicate preferences regarding web scraping and AI training.

While legal compliance cannot be reduced to technical signals alone, such mechanisms could play an important supporting role.

From a practical perspective, many disputes arise because there is currently no consistent and widely adopted way for website operators, content creators and AI developers to communicate with one another.

The development of common standards could help improve transparency, support accountability and reduce uncertainty for all parties involved. I therefore welcome any efforts to encourage further work in this area.

Conclusion

Overall, I strongly welcome these Guidelines and the EDPB's efforts to provide greater clarity in an area that is rapidly evolving.

My principal suggestion is that the final version should place even greater emphasis on practical implementation. Many organisations understand the legal principles. What they need most is clearer guidance on how those principles should be applied in real-world situations involving modern AI systems, complex supply chains and publicly available online information.

Further clarification regarding reasonable expectations, downstream users, transparency obligations and practical implementation measures would, in my view, significantly increase the practical value of the Guidelines while continuing to uphold the high standards of protection expected under EU data protection law.

Thank you for the opportunity to provide feedback.