

CCIA Europe Response to EDPB Consultation

Improving the EDPB guidelines on web scraping in the context of generative AI

October 2026

The Computer & Communications Industry Association (CCIA Europe) welcomes the opportunity to submit feedback on the European Data Protection Board's (EDPB) draft guidelines 03/2026 on web scraping in the context of generative AI. We respectfully share the following nine recommendations with a view to improving these guidelines.

I. Anchoring the guidelines in technical realities to promote legal certainty and coherence

It is essential to ensure that data protection enforcement is aligned with the AI Act and the Digital Omnibus, while reflecting the technical realities of AI model training.

Recommendations:

1. Align data minimisation with the AI Act's data and data governance mandates
2. Recognise the transformative mechanics of AI model training
3. Ensure coherence with the Digital Omnibus negotiations

II. Delivering meaningful transparency through proportionate safeguards

The guidelines should help establish a balanced, risk-based framework for transparency and compliance that protects individual rights while replacing unworkable requirements with effective oversight.

Recommendations:

4. Streamline transparency requirements and clarify disproportionate effort exception
5. Adapt data accuracy and mapping requirements to AI model training realities
6. Focus safeguards on protecting individuals' personal data

III. Ensuring compliance standards remain workable

Regulatory expectations must be grounded in technical feasibility across sensitive data, web protocols, and distinct crawling architectures to ensure rules remain workable in practice.

Recommendations:

7. Adopt a workable standard for incidental sensitive data
8. Treat web protocols as contextual signals rather than legal mandates
9. Distinguish AI training scraping from user-directed retrieval and account for relevant security exceptions

Introduction

Providing clear, horizontal guidance on how data protection rules apply to web scraping in the context of generative AI is key to promoting legal certainty across the European Union.

The Computer & Communications Industry Association (CCIA Europe) appreciates the pragmatic approach of the draft guidelines, especially the explicit recognition that legitimate interest constitutes a valid legal basis under the General Data Protection Regulation (GDPR) for scraping data to train AI models, subject to the three-step test enshrined in Article 6(1)(f) GDPR.

The clarification in paragraph 36 that the data minimisation principle does not prohibit training an AI model with large volumes of data is also a welcome step. The application of the Court of Justice of the European Union's (CJEU) GC & Others (C-136/17) search engine precedent to the incidental collection of special category data provides useful certainty, as assessing compliance based on a controller's actual "responsibilities, powers, and capabilities" is the only workable approach for large-scale scraping.

Moreover, the acknowledgement in paragraphs 25-28 that the disproportionate effort exception enshrined in Article 14(5)(b) GDPR applies to large-scale scraping enhances legal certainty, alongside the clarification in paragraph 29 that Article 11 GDPR protects controllers from being forced to acquire additional identifying information solely to fulfil data subject requests related to unstructured scraped data.

However, specific requirements within the draft guidelines require further refinement to avoid regulatory incoherence, unworkable operational demands, and technical impossibilities. CCIA Europe offers nine recommendations structured around three overarching objectives:

- I. Anchoring the guidelines in technical realities to promote legal certainty and coherence
- II. Delivering meaningful transparency through proportionate safeguards
- III. Ensuring compliance standards remain workable

The Association is committed to supporting the EDPB in setting out guidelines that genuinely promote innovation while protecting data subject rights.

I. Anchoring the guidelines in technical realities to promote legal certainty and coherence

It is essential to ensure that data protection enforcement is aligned with the AI Act and the Digital Omnibus, while reflecting the technical realities of AI model training.

1. Align data minimisation with the AI Act's data and data governance mandates

To ensure legal coherence, the guidelines should closely align with the AI Act. In their current form, they directly conflict with certain AI Act provisions. For instance, Article 10(3)

of the AI Act explicitly mandates training data to be “relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose.” This is broadly echoed in Article 9 of the AI Act, which requires risk management and appropriate testing of high-risk AI systems. These provisions are not blanket training-data mandates for every general-purpose AI model, but they illustrate why broad, representative data is important for safe and trustworthy AI. However, the guidelines currently run against these key requirements.

To avoid contradiction with the AI Act and, most importantly, promote safe and trustworthy AI development, the GDPR’s data minimisation principle cannot be interpreted as requiring narrowness or ‘targeted scraping’ when the AI Act mandates breadth and representativeness. It is important to stress that broad and diverse training corpora are predominantly non-personal and include anything from code to scientific literature and factual text. Narrowing collection only reduces safety, quality, and representativeness, without substantially reducing the amount of personal data collected. The guidelines should therefore recognise that broad, untargeted web crawling can satisfy the necessity condition under Article 6(1)(f) GDPR for general-purpose AI models.

The same distinction matters for the necessity assessment in paragraph 52. Synthetic and licensed data can complement publicly available web data but should not be presumed to provide an equally effective replacement for real-world linguistic, cultural and factual diversity. The models that generate synthetic data themselves depend on real-world training data. The guidelines should recognise the legitimate interest in developing safe, capable general-purpose AI models, including models made available for research and downstream use. The guidelines should allow the necessity of processing to be assessed against those objectives without requiring every future application to be identified at collection.

2. Recognise the transformative mechanics of AI model training

The AI model training process relies on the extraction of statistical patterns, transforming raw input data into generalised representation. The resulting model weights are not a copy or structured database of the training data and should not be presumed to constitute personal data merely because personal data was present at the input stage. This needs to be clarified, as the current guidelines tend to treat this process as resulting in persistent data storage, which is not the case. A residual risk, known as ‘memorisation’, is a marginal technical imperfection that should be assessed in its actual technical context and mitigated through safeguards. It is not evidence that personal data is stored inside the model. The assessment should distinguish the training dataset, the model parameters, and the deployed system’s outputs.

Furthermore, Recital (26) GDPR requires a contextual identifiability assessment that considers all means reasonably likely to be used for identification, from the perspective of the controller or recipient. Isolated personal references or names within a multi-trillion-token training corpus are not personal data by default, as individuals are not identifiable to the model developer without disproportionate effort. The guidelines should apply the GDPR’s own identifiability thresholds rather than assuming all web content that may contain names or images constitutes personal data. As such, the stated GDPR requirements fall away for most untargeted web crawlers.

3. Ensure coherence with the Digital Omnibus negotiations

The EDPB guidelines on web scraping in the context of generative AI are being drafted in parallel with the ongoing legislative negotiations on the Digital Omnibus proposal, which aims to establish horizontal, harmonised rules for AI development, including incidental special category data and legitimate interest.

For legal coherence, the EDPB must ensure that its guidelines remain sufficiently flexible, proportionate, and risk-based to avoid any potential contradiction with the final Digital Omnibus text. The Board should also afford stakeholders a formal opportunity to make further submissions once the Digital Omnibus legislative process has formally concluded. The guidelines should further refer to Article 10 of the AI Act, which permits the processing of special category data for the purposes of ensuring bias detection and mitigation.

II. Delivering meaningful transparency through proportionate safeguards

The guidelines should help establish a balanced, risk-based framework for transparency and compliance that protects individual rights while replacing unworkable requirements with effective oversight.

4. Streamline transparency requirements and clarify disproportionate effort exception

While the guidelines correctly acknowledge that the disproportionate effort exception provided for in Article 14(5)(b) GDPR applies in the context of large-scale web scraping, paragraph 27 of the draft guidelines introduces legal ambiguity by stating that it should not be relied upon “routinely”. The application of Article 14(5)(b) should depend solely on objective factual conditions – such as scale, available contact information, and the steps needed to identify and reach individuals – rather than on the frequency of a controller’s operations. The guidelines should clarify that the exception applies when those conditions are met, including for individuals mentioned in content posted by others, and remove the reference to “routinely”. Controllers should provide meaningful public information and appropriate safeguards when relying on the exception.

Moreover, expecting controllers to publish searchable lists of granular domain names, specific URLs, and precise collection dates under paragraphs 31 is technically and operationally impossible for vast training corpora. It would also expose proprietary trade secrets, run counter to the calibrated training data summary framework under Article 53(1)(d) AI Act, and expose scraping infrastructure to security vulnerabilities such as targeted data poisoning and extraction attacks. Paragraph 30’s suggestion to indicate the types of sources excluded from scraping and the reasons for the exclusion go beyond Article 14 GDPR and would clutter privacy notices with irrelevant negative disclosures.

Detailing scraped sources at a category- or domain-type level, alongside clear rights-exercise channels, provides meaningful transparency without imposing unworkable operational burdens and should be the appropriate standard. Similarly, the suggestion in example 2 to exclude pages where individuals are directly identifiable is not scalable or feasible across billions of web pages. Providing individuals with a list thousands of pages

long and informing them that their data may or may not be included would increase confusion, not transparency.

Finally, the suggestion in paragraph 34 to make Data Protection Impact Assessments (DPIAs) publicly available introduces serious confidentiality and security risks by threatening business and trade secrets, alongside AI system security and the confidentiality of regulatory discussions between controllers and their Lead Supervisory Authority under Articles 35 and 36 GDPR. This measure goes beyond the GDPR, which does not require the publication of DPIAs. These assessments should remain confidential and available only to data protection authorities upon request, while public transparency is delivered through established public-facing channels such as privacy policies and supplementary AI notices.

5. Adapt data accuracy and mapping requirements to AI model training realities

The recommendation in paragraph 42 of the guidelines to scrape from reliable sources such as government registries considerably limits data ingestion and will unavoidably lead to heavily biased and unrepresentative AI models. First, spot-checking cannot provide a high level of confidence in the accuracy of the entire dataset. Second, the vast majority of the information will be difficult to validate, in particular information about non-public figures.

Fact-checking information about private individuals for the sole purpose of validating the training data also intrudes on individuals' privacy. Overall, information about non-public figures on the internet is limited and therefore highly unlikely to result in a risk of regurgitation or memorisation. Moreover, the suggestion that controllers timestamp scraped data to track freshness and conduct pre-training factual verification audits on random samples is incompatible with large-scale ingestion. CCIA recommends removing unworkable expectations for granular timestamping and pre-training verification audits, while broadening the concept of 'reliable sources' to reflect the necessity of diverse, general internet data to mitigate AI model bias. Accuracy safeguards should focus on outputs that concern identifiable individuals, including correction and user-feedback mechanisms.

Furthermore, paragraph 37 of the guidelines calls for a dedicated "data mapping and inventory process" prior to scraping, creating an unnecessary administrative burden that duplicates the standard Records of Processing Activities (ROPA) under Article 30 GDPR. The EDPB should explicitly confirm that maintaining an up-to-date ROPA, alongside appropriate dataset documentation, satisfies the data mapping expectations of paragraph 37 without requiring a separate item-level inventory.

Paragraph 54 should distinguish AI model training, whose purpose is statistical learning rather than profiling individuals, from targeted surveillance and downstream misuse. Available information relating to individuals on the internet is scarce, which considerably reduces the potential risks of memorisation. The suggestion that web scraping for the purpose of AI training could create a sense of mass surveillance for data subjects should be removed. Any assessment of a chilling effect should be grounded in the realistic likelihood and severity of an impact attributable to the processing, taking account of the public context of the sources, the model's architecture, and safeguards against memorisation and

harmful outputs. Public accessibility is relevant to reasonable expectations but should not be equated with consent or a waiver of data subject rights.

6. Focus safeguards on protecting individuals' personal data

Paragraph 65 of the guidelines asserts that mitigating measures in the balancing test “should not be confused with measures that the controller is legally required to adopt anyway to ensure compliance with the GDPR.” The EDPB cites its guidelines 1/2024 in this context.

As it stands, the guidelines go beyond GDPR requirements and objectives without justification. If a technical or organisational safeguard – such as data minimisation, pseudonymisation, or strict security controls – materially reduces real-world privacy risks to individuals, it should carry full weight in the Article 6(1)(f) balancing test even if it also fulfils another statutory GDPR requirement. We encourage the EDPB to revise paragraph 65 to allow controllers to take full account of all risk-reducing measures in their Legitimate Interest Assessment.

The draft guidelines consider only data subjects' interests. For the analysis of the balancing test to be valid, they must also consider the interests of data controllers and third parties. In the context of an epochal technology like AI, the controllers' interests in developing state-of-the-art technology, as well as customers' interests in using it to greatly improve their lives, are both significant factors in the balancing test.

III. Ensuring compliance standards remain workable

Regulatory expectations must be grounded in technical feasibility across sensitive data, web protocols, and distinct crawling architectures to ensure rules remain workable in practice.

7. Adopt a workable standard for incidental sensitive data

Applying the CJEU's GC & Others (C-136/17) framework to incidental special category data collection is pragmatic, but the conditions set out in paragraph 74 remain absolute, overly rigid, and disproportionate. Paragraph 74(ii) requires that incidentally collected special category data be “deleted from datasets immediately”. Scanning the entire internet to ensure the absolute absence of special category data is disproportionate and technically challenging given the current state of the art. Unlike structured identifiers such as phone numbers that follow predictable patterns, special category data (such as political opinions or health status) is inherently contextual, meaning that real-time pre-filtering generates high false-positive rates. Meanwhile, collected training text is rapidly fragmented and distributed across computing systems, making instant extraction operationally unworkable.

Paragraph 73(a) should clarify that ‘relevant similarities’ with search engines concern the structural characteristics of the processing, i.e. the automated, large-scale collection from publicly available sources, without individual selection or review and without the ability to predetermine every data category encountered. It should not depend on whether the downstream product returns search results or generates responses.

The guidelines impose an unworkable double compliance burden by requiring the perfect exclusion of special category data at the point of ingestion while simultaneously requiring state-of-the-art content filters during live deployment to prevent the model from providing special category data at the output stage. Instead of requiring technological perfection at both the input and output stages, the guidelines should promote a holistic approach combining reasonable input-phase safeguards and output-phase filtering. Such an approach would provide a robust risk-mitigation framework.

In particular, paragraph's 74(4) expectation to 'constantly monitor' outputs should be replaced with reasonable, risk-proportionate monitoring of systems a controller actually deploys. A provider releasing open model weights cannot observe or control every third-party deployment. Its safeguards should focus on pre-release evaluation, appropriate deployment guidance and processes to respond to reported issues, while deployment-stage monitoring follows the controller operating the system. Paragraph 77 should likewise require periodic, document review of relevant technological developments, not continuous surveillance of all research or immediate adoption of unproven techniques.

The guidelines should also distinguish between swift deletion from pre-training databases, which is feasible, and post-training deletion, which requires expensive retraining or unlearning.

8. Treat web protocols as contextual signals rather than legal mandates

The draft guidelines risk treating robots.txt and ai.txt as binding obligations for the purpose of data minimisation under Article 5(1)(c) GDPR, which conflates the preferences expressed by website operators with the question of whether personal data is necessary for the controller's purpose. Robots.txt is an established protocol for expressing web crawling directives but has no defined legal status under EU law. It does not constitute a GDPR consent or objection mechanism. The same applies to ai.txt, which has no established standards body and varies considerably in format. While these signals are relevant for the balancing test, they do not constitute an absolute bar to collection. The guidelines should treat them as one factor among many, rather than a determinative factor. The guidelines should move this discussion from paragraph 37 to the reasonable expectations assessment in paragraphs 62-64 and distinguish these signals from authentication and other access controls.

Furthermore, paragraph 66(f) suggests that controllers create an opt-out list to allow "data subjects to object to the collection of their data on certain websites or online platforms by providing information that identifies them on those websites", potentially exposing data subjects to even greater privacy risks. This measure runs counter the privacy-protective approach recognised in Article 11 GDPR and places an impossible identification and administrative burden on controllers. The guidelines should instead recognise accessible rights channels, proportionate verification, existing machine-readable website-operator signals and output-level controls, without presenting robots.txt as a substitute for individuals' GDPR rights.

Similarly, paragraph 66(c)'s suggestion to exclude data on freely accessible websites where individuals cannot be considered "reasonably aware of the accessibility" introduces an unworkable subjective inquiry that belongs to the primary website publisher under Article 13 GDPR. For third-party crawlers, accessibility should be assessed against objective,

technically verifiable criteria, such as whether content is publicly accessible online without authentication barriers or technical access controls. More generally, copyright compliance should be clearly distinguished from data protection enforcement and should not form part of it. Example 7 should therefore remove the copyright-clearance conditions from its GDPR balancing example.

9. Distinguish AI training scraping from user-directed retrieval and account for relevant security exceptions

The guidelines, which focus on large-scale, automated, pre-training collection, should explicitly exclude user-directed, inference-time data retrieval – such as agentic browsing, tool use, or retrieval-augmented generation (RAG).

User-directed retrieval is entirely different from large-scale scraping for AI model training: it is targeted, user-directed, transient, and analogous to web browsing by a user. Moreover, the appropriate legal basis for user-directed retrieval is Article 6(1)(B) GDPR, as it relates to contract performance and user instructions. The guidelines should therefore explicitly exclude this category from their scope of application. For the specific use cases in which data obtained through user-directed retrieval subsequently is used for model improvement, the guidelines should address this process separately and apply measures that are proportionate to the nature and scale of the reuse.

Last but not least, the guidelines should fully account for relevant security and integrity exceptions. Certain crawlers operate only for the purpose of maintaining integrity and security, including those that detect malicious content, verify link safety, and support anti-fraud investigations. Such crawlers are exempted from certain data collection restrictions, including robots.txt instructions, as they would otherwise compromise companies' ability to protect users and detect threats.

Data collected by these specific crawlers is not used for AI model training purposes and is effectively ring-fenced, in line with the principle of purpose limitation. The EDPB itself recognises that “detecting fraudulent content” and “improving threat detection” qualify as legitimate interest for scraping. The guidelines should recognise these exceptions for controllers that maintain technical and organisational barriers preventing the reuse of data crawled for security and integrity purposes, instead of treating all crawling activities bypassing the robots.txt protocol as a negative factor.

Conclusion

To achieve greater legal certainty and coherence, the EDPB guidelines on web scraping in the context of generative AI would benefit from further improvements. They need to be anchored in technical realities, deliver meaningful transparency through proportionate safeguards, and ensure compliance standards remain workable in practice. CCIA Europe is committed to supporting the EDPB in setting out guidelines that genuinely promote innovation while protecting the rights of data subjects.

About CCIA Europe

The Computer & Communications Industry Association (CCIA) is an international, not-for-profit association representing a broad cross section of computer, communications, and internet industry firms.

As an advocate for a thriving European digital economy, CCIA Europe has been actively contributing to EU policy making since 2009. CCIA's Brussels-based team seeks to improve understanding of our industry and share the tech sector's collective expertise, with a view to fostering balanced and well-informed policy making in Europe.

Visit ccianet.eu, x.com/CCIAEurope, or linkedin.com/showcase/cciaeurope to learn more.

For more information, please contact:

CCIA Europe's Head of Communications, Kasper Peters: kpeters@ccianet.org