ITI Promoting Innovation Worldwide

ITI Feedback on the Guidelines 03/2026 on web scraping in the context of generative AI

30 October 2026

The Information Technology Industry Council (ITI) is the premier voice, advocate, and thought leader for the global information and communication technology (ICT) industry. Our member companies include the world's leading innovation companies, with headquarters worldwide and value chains distributed around the globe.

I. Introduction

ITI welcomes the EDPB's efforts to clarify how the GDPR applies to web scraping in the context of generative AI. The draft Guidelines recognize several practical realities of large-scale model development, including the possible use of legitimate interests, the need for substantial training datasets, and the limits of individual notice. Those clarifications should be preserved. Other passages risk turning illustrative safeguards into a demanding compliance framework that is difficult to implement and may diminish the diversity and quality of training data. In the spirit of the EDPB's Helsinki Statement, ITI recommends several changes so that the final guidance reflects a technology-neutral, risk-based approach that protects individuals while giving controllers workable ways to demonstrate compliance.

II. Areas of support

2.1 Legitimate interests and data minimization

ITI welcomes the recognition that Article 6(1)(f) GDPR may provide a lawful basis for web scraping for AI training, subject to a case-specific three-step assessment. We also welcome paragraph 36's clarification that data minimization does not itself rule out processing large volumes of data. ITI reiterates that a broad, diverse collection of training data can be necessary for model performance, representativeness, and safety. This is further underlined by Article 10 of the AI Act, which mandates that training data should be relevant, sufficiently representative, and complete. Whether the size of the training dataset is necessary should be assessed against a defined training purpose and the available safeguards, rather than volume alone.

2.2 Incidental Special Category Data (SCD)

The draft reference to the Court of Justice's approach in *GC and Others (C-136/17)* offers a useful basis for assessing incidental special category data considering a controller's responsibilities, powers and capabilities. This approach should recognize what a controller

United States
700 K Street NW, Suite 600
Washington, D.C. 20001, USA
+1 202-737-8888

Europe
Rue Froissart 95, 1040
Brussels, Belgium
+32 (0)2 321 10 90

info@itic.org
www.itic.org
@iti_techtweets

can reasonably identify and mitigate in very large, changing datasets, while maintaining meaningful safeguards for data subjects.

2.3 Transparency and identification of data subjects

ITI supports the recognition in paragraphs 25–28 that individual notice may involve disproportionate effort within Article 14(5)(b) GDPR. Paragraph 29's reference to Article 11 is equally important: a controller should not have to acquire additional identifying information solely to link individuals to unstructured training data to respond to Data Subject Requests (DSRs). Controllers should still provide meaningful public information and accessible channels for rights requests where identification is possible.

III. Recommendations for revision

3.1 Apply the GDPR threshold for personal data in context

The Guidelines should avoid assuming that every item of publicly available web content or every model weight is personal data. Identifiability requires a contextual assessment under Recital 26, including the means reasonably likely to be used by the controller or another person. The same analysis should be applied to training data and to a trained model, with due attention to memorization and extraction risks. This is not a categorical claim that model weights can never contain personal data. Identification in this context also requires more than distinguishability within a dataset: it requires that the controller be able to single out an individual and, as a consequence, treat that person differently from others, which is the risk the GDPR is designed to address (Recital 75). The EDPB should complete its guidance by clarifying how controllers can document a proportionate assessment rather than requiring impossible item-by-item identification across a corpus.

3.2 Make transparency obligations workable

The availability of the Article 14(5)(b) exception should turn on its statutory conditions, including the actual scale and feasibility of individual notice, rather than whether reliance on it is described as "routine". Example 2 should not imply that controllers must reliably exclude every page on which a person is directly identifiable before collecting at scale. Such a rule would be difficult to operationalize and would conflate identifiability with whether individual notice is feasible.

However, by recommending disclosure of "precise" source information and requiring privacy notices to list the types of sources excluded from scraping and the reasons for their exclusion (paragraph 30), the draft goes beyond the requirements of Article 14(2)(f) of the GDPR and runs counter to the AI Act. Article 14(2)(f) requires the source "from which . . . the personal data originate, and if applicable, whether it came from publicly accessible sources." It does not require an enumerated index or negative disclosures of uncollected data that would clutter privacy notices and confuse users. A searchable, per-URL disclosure of a

training corpus itself can enable re-identification, since it could allow bad actors to link individuals to the specific pages from which their data was obtained and use that information for identity theft, targeted data poisoning, or other abuse, while exposing proprietary trade secrets regarding dataset composition, resulting in the opposite of the safeguards' intent. Granular disclosure of training data composition also provides a roadmap for adversarial attacks such as membership inference, data extraction, and targeted data poisoning. The final Guidelines should:

(i) state clearly that the searchable-URL practice is illustrative and not a *de facto* requirement; and

(ii) recognize category-level source descriptions as sufficient where enumeration would create the risks above and/or where precise URL lists and page-level dates cannot reasonably be produced.

Controllers should explain what information they can provide and how people may exercise their rights. Article 14 does not require disclosure of granular crawler characteristics or individual data-provider contact details; categories and general sources of data should ordinarily suffice, particularly where greater granularity is infeasible, creates security or re-identification risks, or duplicates AI Act disclosures.

The suggestion in paragraph 34 to publish Data Protection Impact Assessments (DPIAs) should be removed: Article 35 does not impose general publication, and DPIAs—even in redacted or summary form—may reveal confidential information, proprietary trade secrets, and security-sensitive details, while undermining the confidentiality of regulatory discussions with the Lead Supervisory Authority (LSA). Publication would also expose vulnerabilities that a controller has assessed but not yet fully mitigated and may affect legally privileged analysis; Articles 35 and 36(3) confirm that the GDPR's design is availability to supervisory authorities rather than publication. A more reasonable recommendation would be to make DPIAs available to supervisory authorities upon request, while relying on established external surfaces such as Privacy Policies and supplemental AI notices to inform the public.

The final Guidelines should expressly recognize the AI Act's transparency framework for GPAI models and clarify how GDPR transparency obligations interact with the Commission's mandatory Template for the Public Summary of Training Content. Recommendations to disclose searchable lists of domains or URLs should not result in a parallel or more granular training-data disclosure regime that undermines the calibrated approach adopted under Article 53(1)(d) AI Act. ITI would also strongly encourage the EDPB to consider promoting the adoption of international standards and best practices around transparency, including those existing and currently being explored in ISO/IEC SC 42.

3.3 Assess legitimate interests without privileging narrow crawling

The balancing test should recognize that broad crawling can be necessary for general-purpose models where diverse data supports language coverage, representativeness, quality, and safety. Neither targeted crawling nor an illustrative collection of preferred

safeguards should become a de facto prerequisite. Furthermore, the Guidelines' structural preference for 'targeted scraping' over broad, general web crawling misinterprets the technical architecture of foundational models, which inherently require diverse, untargeted datasets to ensure representativeness and mitigate algorithmic bias. The EDPB should provide realistic examples in which a documented combination of source exclusions, personal data reduction, retention limits, output controls, transparency and effective objection channels supports the balancing test. Measures already required by the GDPR can genuinely reduce the impact on individuals and should be considered in the factual assessment even where they do not, by themselves, establish a lawful basis (paragraph 65). A documented legitimate interests assessment (LIA) may cover a defined class of materially similar scraping and training activities. It should be refreshed where the purpose, source categories, safeguards, retention period, or risk profile materially changes. A refresh should not be required merely because a model is retrained, versioned, or deployed.

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

While acknowledging that publicly accessible material can still give rise to privacy concerns, the guidelines should acknowledge the purpose of model training is not related to profiling individuals and thus Paragraph 54 should distinguish AI training from surveillance or profiling of named individuals. The draft cites no empirical evidence that AI training data collection has caused measurable self-censorship, and chilling-effect reasoning derives from contexts in which individuals are monitored by an identifiable observer able to act on what is observed. Example 7 should keep copyright rules and the GDPR balancing exercise analytically distinct. The copyright condition should therefore be removed from Example 7: the EDPB has no mandate to interpret or condition GDPR compliance upon Directive 2019/790. The requirement that data subjects have “manifestly made the content public” should likewise be replaced by a reference to publicly accessible sources, as it imports a threshold from Article 9(2)(e) into the Article 6(1)(f) balancing test, where public accessibility is a relevant factor rather than a precondition. Compliance with one regime does not automatically determine compliance with the other.

The draft appropriately identifies synthetic data, anonymization, and pseudonymization as possible safeguards. However, if these are interpreted as default requirements, the guidance could undermine legitimate AI training where fully synthetic data or full anonymization would materially degrade the purpose or model quality. The final Guidelines should clarify that synthetic data, anonymization, and pseudonymization controls are proportional tools, not mandatory defaults. The Guidelines should recognize that synthetic data currently lacks quality as training data.

3.4 Set reasonable crawling exemptions without considering them as conclusive

While we agree with the general principle that websites which clearly oppose the scraping of their content through the use of established machine-readable signals such as robots.txt should be excluded from automatic data collection, these should not be described as automatic legal prohibitions under data minimization in the context of GDPR. Such signals are better assessed within the balancing test: robots.txt was designed as crawling etiquette for search engines, has no defined status under EU law, and is frequently misconfigured or over-broad. The draft should not treat ai.txt as an equivalent or mandatory standard: its format and adoption remain unsettled.

The draft suggests that a controller may mitigate risk in the balancing test under Article 6(1)(f) of the GDPR by implementing a prior objection mechanism or creating new opt-out lists at the individual level. These measures are not yet feasible at web scale. Instead, existing mechanisms operating at the domain or URL level, such as robots.txt, could be recommended. Creating and maintaining a person-level register would require controllers to acquire and retain identifiers, creating additional processing, security, and Article 11 concerns. The draft's treatment of an unavailable mechanism as an expected safeguard could penalize controllers for the absence of shared infrastructure and create obligations that conflict with minimization, accuracy, and proportionality principles.

Paragraph 66(f) should avoid implying that every controller must maintain a global list identifying people across websites. Rights and objections should be handled through proportionate channels that do not require individuals to disclose additional identifiers unnecessarily. Accordingly, the expectation for global opt-out lists in Paragraph 66(f) should be removed in favor of promoting standard, machine-readable protocols like robots.txt, which serve as a proven, scalable mechanism for managing automated data collection. 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 (whether content is publicly accessible online without authentication barriers or technical access controls).

3.5 Clarify which crawling activities the Guidelines address

The Guidelines should distinguish bulk collection for model training from user-directed, real-time retrieval for an immediate response, such as fetching a URL supplied by a user or retrieval-augmented generation. Purpose, controller role, retention and lawful basis may differ, so the training-focused examples should not automatically be transposed to these activities. Article 6(1)(b) may be relevant to some user-requested services where its necessity conditions are met; it should not be presumed to apply in every case. Any later reuse of retrieved material for training or improvement requires its own purpose and lawful-basis assessment.

Crawlers used to detect fraud, malicious content, or other security threats should likewise be assessed according to their distinct purpose and risks. The Guidelines should recognize

technical and organizational separation that prevents security-collected data from being repurposed for AI training without a separate assessment. The mere operation of a security crawler should not be treated as adverse evidence in the balancing test for an unrelated training activity.

3.6 Align accuracy and minimization with model training

The accuracy principle should be applied with regard to the purposes and risks of model training. Paragraph 42's preference for "reliable sources" should not be read as a requirement to confine training to official registers or curated sources; doing so could reduce coverage and representativeness. Imposing traditional static database accuracy expectations onto AI ingestion is technically unfeasible; limiting training data exclusively to highly curated or 'official' sources inevitably introduces systematic bias and degrades model utility.

Granular timestamps and factual verification of random source items may be ineffective at corpus scale and may require intrusive additional investigation about people. Consequently, the Guidelines should omit expectations for granular timestamping and pre-training factual verification audits of random samples, which are incompatible with the technical realities of web-scale ingestion. The EDPB should instead recognize proportionate source-quality controls, deduplication, dataset documentation, evaluation and correction measures matched to the model's intended use and the likelihood of harmful inaccurate outputs. Paragraph 37 should clarify that existing records of processing activities and related dataset documentation may fulfill its data-mapping expectations where they contain the relevant information. It should not imply a separate, duplicative inventory.

The Guidelines should also read GDPR minimization coherently with relevant AI Act requirements on data quality and representativeness, without suggesting that those requirements displace the GDPR or apply identically to every model. The current version may not give sufficient weight to effective downstream controls when some unnecessary or sensitive data can be identified only after extraction (para. 36-38). The final Guidelines should clarify that controllers may satisfy data-minimization obligations through a documented combination of safeguards across the AI-development lifecycle. These may include excluding high-risk sources and applying collection filters where feasible, followed by post-collection detection, deletion or pseudonymization of unnecessary or sensitive data and, where deletion is technically impossible or disproportionate, through access, retention, model-training, testing, and output controls designed to prevent that data from being revealed or inferred. Where a data subject objects under Article 21(1) in relation to a trained model, output-level controls should be recognized as a proportionate and technically appropriate remedy, by analogy with Google Spain (C-131/12), where the Court required intervention at the point of access rather than deletion of the underlying source.

The Guidelines should make clear that an organization's responsibilities must reflect the information reasonably available to it and its practical ability to verify that information, particularly where it did not collect or prepare the underlying data. As suggested in the

draft, a controller preparing training data should take reasonable steps to use reliable sources, record when the data was collected, check the data for obvious errors, and correct known problems. An organization using the trained model should consider additional safeguards related to the model's output based on how the AI system will be used and the harm an incorrect answer could cause. However, the Guidelines should not suggest that the controller preparing the training data must guarantee that a general-purpose model will always produce correct personal information. Importantly, giving each organization responsibility for the risks to accuracy that it can actually control is consistent with the AI Act's broader risk-based approach.

3.7 Make safeguards for incidental special category data (SCD) proportionate

Paragraph 74 should call for reasonable, documented measures appropriate to the risk and the available technology, rather than an absolute guarantee that all incidental SCD is detected and deleted immediately on ingestion. Unlike structured identifiers such as phone numbers, special category data (such as political opinions or health status) is inherently contextual, meaning real-time pre-filtering across billions of pages causes high false-positive rates, while scraped text is rapidly tokenized and fragmented across computing clusters, making instant extraction operationally unworkable.

It is also important that the Guidelines do not presume that SCD must always be excluded from training databases. Subject to an appropriate Article 6 lawful basis and Article 9 condition, limited incidental processing may sometimes be relevant to representativeness, bias mitigation, safety and culturally appropriate performance. The focus should be a documented, risk-based set of safeguards rather than a blanket exclusion rule or a presumption that any output touching SCD is inherently inappropriate.

A controller should combine appropriate source controls, detection and removal where feasible, restrictions on retention and access, testing for memorization, and safeguards at deployment. Output filters can complement ingestion safeguards; neither layer alone should be assumed sufficient in every case. The effectiveness of the overall measures should be evaluated and improved as risks and techniques evolve. The final Guidelines should avoid imposing an unworkable double compliance burden that demands both perfect exclusion of SCD at ingestion and state-of-the-art content filtering during deployment, acknowledging instead that a risk-based combination of these measures constitutes sufficient mitigation. The Guidelines should also distinguish between swift deletion from a pre-training database, which is feasible, and post-training deletion, which requires expensive retraining or unlearning.

3.8 Clarify responsibilities for recipients

The draft correctly distinguishes between the entity that originally scrapes data and a later AI developer that obtains and reuses the resulting dataset. It states that each entity is, in principle, responsible for its own processing because each determines the purposes and

essential means of that processing. For transparency, the draft recommends that a downstream developer ("recipient") identify the upstream controller, provide a means of contacting it, link to its website, and explain clearly the conditions under which the personal data was collected.

However, the draft does not clearly specify which checks recipients should carry out before using the relevant data. These could cover the data's source and collection methods, the applicable legal basis, any measures taken by websites to prevent scraping, the handling of sensitive data, transparency and the exercise of individuals' rights, retention periods, data security, and the reliability of the supplier's claims (see paras. 16–20). The final Guidelines should clarify the reasonable, risk-based checks expected of a recipient before it uses a scraped dataset. The standard should require proportionate verification of the supplier and dataset, not reconstruction of every upstream URL or collection event.

3.9 Preserve coherence with legislative developments

The Guidelines should remain within the GDPR and avoid presenting optional practices as new legal duties. They should be revisited, where necessary, considering the final text of any Data and Privacy Digital Omnibus measures addressing AI development, incidental data and legitimate interests, and the EDPB should afford stakeholders a formal opportunity to make further submissions once the Digital Omnibus legislative process concludes. This will help ensure coherent interpretation without prejudging ongoing legislative negotiations. The EDPB must maintain sufficient flexibility in Guidelines 03/2026 to ensure they do not run counter to the harmonized, horizontal rules for AI development currently being finalized in the Digital Omnibus political negotiations.