Ethical Web Data Collection Initiative
Response to the European Data Protection Board's Draft Guidelines 03/2026 on Web Scraping in the Context of Generative AI

Executive Summary & Introduction

The Internet Infrastructure Coalition (i2Coalition) and the Ethical Web Data Collection Initiative (EWDCI) welcome the opportunity to respond to the European Data Protection Board's (EDPB) draft Guidelines 03/2026 on Web Scraping in the Context of Generative AI. We commend the EDPB for undertaking a comprehensive effort to clarify how existing General Data Protection Regulation (GDPR) requirements apply to the automated collection of publicly available information for artificial intelligence models.

The i2Coalition represents the foundational infrastructure providers of the global Internet, including web hosting, cloud computing, managed services, data centers, domain registries and registrars, and cybersecurity providers. EWDCI, an initiative of the i2Coalition, is an international industry-led consortium dedicated to strengthening public trust, establishing and implementing ethical web data collection practices, and promoting responsible data aggregation.

Open automated data collection and interoperable standards have long supported core web functions, including search indexing, cybersecurity threat detection, digital preservation, accessibility services, and academic research. As the regulatory framework for generative AI develops, it is crucial that guidelines maintain openness, security, and innovation without creating unintended regulatory barriers for legitimate, well-established, commonly used low-risk Internet operations.

This submission offers constructive, policy-focused recommendations to ensure that the final Guidelines remain technology-neutral, operationally feasible, and legally sound while upholding high data protection standards across the European Union.

Recommendation 1: Clearly Distinguish Between Purposes of Automated Data Collection

Context and Feedback

Automated data collection on the Internet encompasses a wide variety of distinct operational purposes. Organizations routinely gather publicly available web information for search engine indexing, threat intelligence, vulnerability monitoring, research, digital archives, market

intelligence, price comparison, and generative AI training. While these activities frequently utilize similar technical mechanisms, they serve vastly different public functions and carry distinct risk profiles. Lawful access to publicly available information, including through automated collection, supports a wide range of digital services and contributes to Europe's digital economy and technological development.

The Guidelines correctly note that information being publicly accessible does not remove it from GDPR protections. We also welcome the Guidelines' recognition that training generative AI models requires vast amounts of data from diverse sources (paragraph 2), and that the principle of data minimisation does not prohibit training an AI model with large volumes of data (paragraph 36). While these Guidelines "aim to clarify the legal and technical implications of web scraping for generative artificial intelligence (AI) development under the GDPR," they risk being applied to all automated data collection. However, treating all automated data collection under a uniform, AI-centric regulatory umbrella risks inadvertently burdening non-AI, essential web services with constraints designed specifically for generative model training. Web scraping is a delivery mechanism, functionally comparable to obtaining the same data through an API, a feed, or manual collection. Where personal data is collected and processed in full compliance with GDPR principles, including an appropriate lawful basis, purpose limitation, data minimisation, and transparency, the use of automated collection should not in itself trigger additional regulatory barriers.

The specific risks the Guidelines identify in relation to generative AI arise largely from characteristics particular to model training rather than from automated collection as such. While training datasets can be stored and individual records can be located and removed, the difficulty with erasure and rectification arises once data has been used to train the model. Training datasets also derive their value from breadth, which means data minimisation and filtering mostly happens once the dataset has been assembled. This use and preservation of data is different from other established collection activities. Digital archiving, although it may also collect broadly, preserves web content in its original form and context, often under statutory legal deposit mandates; in this context, individual records remain identifiable and can be located, restricted, or removed where necessary. Security monitoring and threat intelligence focus on specific indicators of risk, such as phishing pages, malware distribution sites, exposed vulnerabilities, or leaked credentials, extracting the signals relevant to detect and respond to those threats; they generally discard the remaining content, and any personal data encountered is largely incidental to the purpose.

Members' operational examples show how public search-result data supports these distinct purposes. Businesses compare prices and availability; SMEs and their agencies monitor their visibility across languages and countries, including how they are represented in AI-generated answers. Researchers, journalists and fact-checkers track public claims, while rightholders identify unauthorised copies of protected works. Monitoring what an AI system says about a business is distinct from collecting data to train that system.[1]

[1]Operational examples supplied by SerpApi to EWDCI on Oct. 28, 2026. These examples describe collection purposes; their lawfulness depends on the specific processing and applicable legal requirements.

EU consumer-protection practice also illustrates the value of automated collection. The Commission's eSurveillance WebCrawler helps enforcement authorities detect online offers of products already notified as dangerous in Safety Gate. That use does not establish a lawful basis for every commercial collection activity, but it demonstrates why automated collection should be assessed by its purpose and effects rather than treated as inherently suspect.[2]

These examples may involve business information or other non-personal material, but can also involve personal data. The distinction must be assessed in the particular case. GDPR Recital 49 recognises strictly necessary and proportionate processing for network and information security as a legitimate interest. A commercial purpose is not automatically excluded from Article 6(1)(f), but necessity and the balancing of individuals' rights remain essential. Journalistic, academic and research activities likewise remain subject to the applicable GDPR provisions and national law. None of these examples creates a categorical exemption from data protection obligations.[3]

Within any given purpose, including AI development, appropriate safeguards will depend on the specific processing concerned. We welcome the Guidelines' recognition that mitigating measures "should be tailored to the circumstances of the case" (paragraph 65). Consistent with Articles 24 and 25 GDPR, the nature, scope, context, and purposes of the processing, together with its likely effects on individuals, should guide which measures are appropriate. Scale and automation are relevant factors, but should not on their own determine what safeguards are required.

We therefore encourage the Guidelines to draw a clear regulatory distinction between automated collection for the purpose of generative AI model training and other established collection activities. Operational safeguards recommended for AI training should be expressly scoped to that purpose, and should not be read as general restrictions on automated web access or on the operation of Internet infrastructure.

Recommendation to the EDPB

- **Validate Web Scraping as a Technical Method:** Explicitly confirm that where personal data is collected and processed in full compliance with GDPR principles (including lawful basis and data minimisation), web scraping as a technical delivery mechanism should not face additional regulatory barriers.
- **Distinguish Operational Purposes:** Draw clear regulatory distinctions between automated collection conducted for generative AI model training versus other

[2]European Commission, Communication COM(2025) 63 final, describing the eSurveillance WebCrawler's support for market-surveillance authorities, https://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2025/0063/COM_COM%282025%290063_EN.pdf; European Commission, eSurveillance – Product Safety, https://esurveillance.ec.europa.eu/.

[3]Regulation (EU) 2016/679, Recital 49 and Articles 6(1)(f), 85 and 89, https://eur-lex.europa.eu/eli/reg/2016/679/oj; CJEU, Koninklijke Nederlandse Lawn Tennisbond, C-621/22, EU:C:2024:857 (Oct. 4, 2024), operative part, https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ%3AC_202500686. NIS2 Article 29 separately provides for voluntary cybersecurity information-sharing arrangements; it does not itself authorise all collection of public web data.

established collection activities such as security monitoring, academic research, and search indexing - ensuring the operational safeguards recommended for AI training models are not interpreted as broad restrictions on general web automated access or infrastructure operations.
- **Support Proportionate, Context-Specific Safeguards:** Confirm that the safeguards appropriate to web scraping should reflect the purposes of the processing, the data involved, and the actual risks to the rights and freedoms of individuals, rather than the use of automated collection alone.

Recommendation 2: Limit the Regulatory Interpretation of Technical Signals and Anti-Scraping Controls

Context and Feedback

Legacy technical mechanisms such as robots.txt, CAPTCHAs, and Web Application Firewalls (WAFs) were originally designed to manage network traffic, prevent Distributed Denial of Service (DDoS) attacks, mitigate fraud, and control server resources. They were not designed to convey the privacy choices of individual data subjects or to distinguish personal data from non-personal web content.

Furthermore, these signals are implemented by website operators—who often hold distinct commercial interests in controlling data access—rather than the individuals whose data appears online. As a result, an operator's technical configuration cannot serve as a reliable proxy for individual data subject intent.

In operational practice, these mechanisms are frequently deployed as blunt, domain-wide configurations:

- **Generic Implementations**: WAFs and CAPTCHAs are routinely provisioned across entire domains via default Web Application Firewall settings, unrelated to specific page content.
- **Unnuanced Directives**: Site operators frequently deploy single-click, domain-wide robots.txt rules without evaluating the content or data on individual pages, banning any crawler outright, and leaving the robots.txt without any actual meaningful context.
- **Dynamic Black-Box Security**: Modern anti-bot tools rely on probabilistic machine learning models analyzing TLS fingerprints, IP reputation, and request timing. These reactive security measures vary by request and visitor, providing no stable, machine-readable declaration about data collection.

These mechanisms also differ materially from one another. A robots.txt directive may be addressed only to a particular crawler or user agent, CAPTCHAs and authentication may protect only specific pages or functions, and newer files such as ai.txt are not yet established standards.

Treating dynamic or site-wide blocking as a definitive expression of GDPR data subject objection creates an unfair imbalance. Sophisticated entities with extensive engineering resources - such as the latest scraping technology, high-reputation residential or mobile IP addresses, browser and fingerprint management - can bypass automated detection flags. Meanwhile well-intentioned, transparent collectors—such as smaller enterprises, academic institutions, and non-profits—are routinely blocked. Interpreting technical blocks as legal objections rewards evasion techniques rather than responsible compliance.

We welcome the Guidelines' recognition that the absence of a robots.txt file does not amount to consent (paragraph 45). By the same reasoning, the presence of such signals should not be treated as an expression of data subjects' choices. We would therefore encourage the EDPB to revisit the final bullet of paragraph 37, which lists the exclusion of websites using robots.txt, ai.txt or CAPTCHAs among the measures controllers should take before collection.

Recommendation to the EDPB

- **Confirm the Limited Weight of Technical Measures:** Confirm that site-wide CAPTCHAs, dynamic WAF blocks, and generic robots.txt entries are not equated with an objection under Article 21 GDPR, and that their weight under paragraph 64 depends on data subjects' awareness and on whether the measure is specific to the use concerned.
- **Clarify the Significance of Different Signals:** If the EDPB retains references to technical signals (paragraphs 37 and 64), clearly define their relevance and distinguish between them according to their purpose, technical reliability, and application to the particular content being collected.

Recommendation 3: Support Open Technical Standards and Interoperability

Context and Feedback

Throughout the history of the Internet, legal frameworks and open technical standards have played complementary roles. Policymakers establish legal rights, obligations, and regulatory objectives, while international standards development organizations (SDOs) define the interoperable technical specifications required for implementation.

Adopting informal or unstandardized technical elements into regulatory mandates creates market fragmentation and technical ambiguity. Explicitly recognizing the division between legal mandates and technical standardization allows Internet technologies to advance without requiring frequent updates to underlying data protection guidance. By supporting open standards, we should also make room within SDOs for smaller organizations and civil society, keep the process transparent, and avoid technical approaches that only the biggest players can realistically implement.

Recommendation to the EDPB

- **Acknowledge the role of Standards Bodies:** Formally recognize the role of established, open Standards Development Organizations (such as the IETF and W3C) in defining (1) machine-readable AI usage preferences, (2) verifiable crawler identification, and (3) interoperable ways to communicate collection preferences - while leaving legal questions (lawful basis, weight in the balancing test, transparency obligations) with the EDPB.
- **Maintain Regulatory Resilience:** Focus regulatory guidance on required data protection outcomes (such as transparency and objection mechanisms) while leaving technical execution to open, interoperable standards.

Recommendation 4: Adopt a Proportionate Tiered Framework for Transparency

Context and Feedback

The draft Guidelines set out the information controllers should make publicly available where informing data subjects individually would be impossible or involve disproportionate effort, including a precise indication of sources and the crawler's characteristics (paragraph 30). They then describe further measures as good practice or as measures controllers may consider: providing the domain names and URLs of scraped pages in a searchable format "where possible" (paragraph 31), information about suppliers of already-scraped datasets (paragraph 32), publishing a DPIA (paragraph 34), and publishing an updated list of scraped websites (paragraph 66(e)). We welcome this distinction. However, even as a good practice, publishing detailed, page-level URL registries of collected personal data would introduce severe privacy and security risks:

- **Aggregation Risks:** A public, searchable registry of exact URLs containing personal data serves as an explicit, consolidated directory of personal data locations online. This directly assists malicious actors in compiling target lists, executing scraping campaigns, and conducting re-identification attacks against data subjects.
- **Secondary Personal Data Processing:** Compiling and hosting a searchable index of pages associated with identifiable individuals constitutes an additional layer of personal data processing, compounding privacy risks.
- **Ineffective Transparency:** Raw lists containing millions of granular URLs provide minimal practical value to individuals attempting to exercise their GDPR rights.

We also welcome the Guidelines' recognition that individual notification may be impossible or involve disproportionate effort when data is scraped at scale and the controller has no means of contacting the individuals concerned, assessed with regard to the dataset as a whole (paragraphs 25 and 28, and Example 2), and that controllers are not required to acquire

additional information to identify data subjects solely for the purpose of complying with the GDPR (paragraph 29). It would be helpful to clarify how this relates to paragraph 27, which states that the exception should not be routinely relied upon outside archiving, research and statistical purposes, and to confirm that controllers may rely on it where the assessment in paragraph 28 supports doing so, as Example 2 illustrates.

The CJEU's search-engine case law provides a relevant analogy for calibrating safeguards to an intermediary's particular role. In GC and Others, the Court considered the search-engine operator's responsibilities, powers and capabilities, including the handling of sensitive-data issues following a data subject's request. The EDPB should similarly consider what a collector of public search data can meaningfully identify and control when assessing appropriate transparency and rights-handling measures. This analogy does not transfer the search-engine rules wholesale to other collectors or displace their obligations under Articles 12–14 GDPR.[4]

Recommendation to the EDPB

Recognise a structured, tiered transparency model as an appropriate way to provide the public information described in paragraph 30, in place of page-level URL publication. Such a model could provide meaningful public insight while safeguarding system integrity and individual privacy:

Disclosure Tier	**Audience**	**Content Scope**
Public Transparency	General Public & Data Subjects	**High-level source categories and registered domain names.** Source categories and domains, including date ranges of data collection periods and a reasoned statement of what is excluded and why.
Public Transparency	General Public & Data Subjects	**Plain-language methodology statement** describing collection behaviour.
Public Transparency	General Public & Data Subjects	**Individual lookup tool**: An optional domain-level query interface allowing individuals to verify if a specific website or domain was indexed.
Regulatory Disclosure	Supervisory Authorities (DPAs)	**Upon formal request**: complete, granular URL-level logs and detailed collection records as well as technical crawler attributes and operational configuration details - provided securely scoped to the inquiry and pseudonymised where possible.

[4]CJEU, GC and Others v CNIL, C-136/17, EU:C:2019:773 (Sept. 24, 2019), paragraphs 37 and 47 and operative part, https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A62017CJ0136. The judgment concerns search engines and de-referencing under Directive 95/46; its role-specific reasoning is invoked here by analogy, not as a general exemption for scraping.

Recommendation 5: Frame Guidance for an Evolving Internet

Context and Feedback

Web data collection architectures are transitioning rapidly from traditional, scheduled web crawlers to autonomous, agentic AI systems. Autonomous AI agents collect information dynamically and in real time across unpredictable web paths to complete user-defined tasks - collecting data in a manner very different from that of conventional crawlers. Measures such as publishing source lists, describing crawler characteristics, and offering pre-collection objection mechanisms presuppose that the data collector knows in advance where it will go. Agentic systems make this assumption increasingly unreliable. It may also be unclear who the data controller is when an agent acts for a user and is built by one organisation that runs on another's infrastructure. Existing GDPR principles already provide the answer: responsibility should follow actual control over the purposes and means of the processing.

Real-time retrieval by an AI assistant provides a concrete example. An application may fetch current search results in response to a user's question and use those sources to ground a cited answer. Such query-time retrieval is operationally distinct from assembling a corpus for model training, even though both may use automated web access. Its sources can be selected dynamically for the particular request, making training-oriented assumptions about advance source lists and pre-collection objection mechanisms a poor fit. This supports our request that the EDPB clarify the scope of these training-focused Guidelines; it does not remove the GDPR obligations applicable to any personal data processed during retrieval.[5]

The example of an increasingly agentic Internet emphasizes the importance of safeguards being framed around outcomes rather than a planned, defined crawl. These technologies and methods of collecting publicly available information will continue to change rapidly. The EDPB shouldn't need to develop a separate set of rules for AI agents. Building guidance around durable principles - grounded in accountability, transparency, proportionality, and responsible processing - rather than assumptions tied to current technologies may help ensure that it continues to provide useful direction as the Internet evolves.

Outcome-focused guidance also benefits from flexibility in how its measures are applied. We welcome the Guidelines' statement that the measures in paragraph 66 "do not constitute an exhaustive list" and that the controller "should make a case-by-case assessment to determine which measures are appropriate." The data minimisation measures in paragraphs 37–38 would benefit from the same clarity. The executive summary presents them as measures that "could be" taken, while paragraph 37 states that the controller "should" take them before collection, and several presuppose a planned, defined crawl. Aligning paragraph 37 with the executive summary would confirm that controllers can select the minimisation measures appropriate to their processing.

---

[5]SerpApi, operational examples supplied to EWDCI on Oct. 28, 2026, example 4 (real-time retrieval to ground AI assistants). The distinction is between retrieval and model training; it does not establish that all retrieval is lawful or outside the GDPR.

Recommendation to the EDPB

- **Adopt Outcome-Focused Neutrality**: Ground guidelines in technology-neutral principles centered on controller accountability, transparency, and data minimisation rather than static crawler assumptions.
- **Confirm Agentic AI is Out-of-Scope:** Confirm that real-time retrieval by agents falls outside these training-focused Guidelines.
- **Confirm Flexibility in Data Minimisation Measures:** Align paragraphs 37–38 with the case-by-case approach of paragraph 66, confirming that the listed measures are options controllers may select according to their processing rather than cumulative requirements.

Conclusion & Next Steps

The i2Coalition and EWDCI appreciate the EDPB's commitment to clarifying data protection standards in the rapidly moving field of generative AI. By addressing the recommendations above, we believe the EDPB can establish a legal framework that robustly protects individual privacy while preserving the open technical infrastructure of the global Internet. We welcome further engagement with the EDPB, national Supervisory Authorities, and industry stakeholders to refine these recommendations into actionable, durable regulatory guidance.