Response to the EDPB draft Guidelines 03/2026 on web scraping in the context of generative AI

Google welcomes the opportunity to provide feedback to the European Data Protection Board (**EDPB**) on its draft Guidelines 03/2026 on web scraping in the context of generative AI (the **Guidelines**). Our response contains some general observations on the Guidelines together with a number of specific points of interpretation that we would be grateful for the EDPB to consider.

Google is deeply committed to the responsible development of AI technologies. We strongly believe that clear, coherent guidance from the EDPB is central to European competitiveness and the successful adoption of such technologies. Protecting individual privacy is foundational to our approach to AI governance. In practice, responsible AI developers deploy comprehensive measures designed to protect personal data throughout the lifecycle of training and deploying models - including honouring machine-readable web publisher controls (such as robots.txt), applying filtering to remove a range of identifying information (such as structured personal identifiers) and harmful content, and maintaining robust output filters. Our comments below are intended to ensure that the Guidelines support these effective, privacy-protective practices while remaining grounded in the technical and operational realities of how modern foundation models are built.

1. Special category data (SCD)

We welcome the EDPB's express recognition that the *GC and Others (C-136/17)* judgment is relevant for the incidental and residual collection of SCD in the context of webscraping for the purposes of AI model training (paragraphs 72 and 73). This approach provides a workable framework for large-scale web-scraping where organisations do not intend to collect SCD, but where completely preventing such collection is technically impossible.

However, while we support the threshold criteria in paragraph 73 of the Guidelines for bringing incidental and residual processing within the scope of GC and Others, we have concerns relating to certain prescriptive technical measures described elsewhere in the Guidelines if they are interpreted as strict prerequisites that the EDPB expects controllers to satisfy in order to rely on the *GC and Others* judgment in this context. In particular, the EDPB's expectations that controllers should apply precise filters for data and websites *prior* to data collection and

ensure the immediate deletion or replacement of any inadvertently collected SCD fail to account for the challenges associated with contextually identifying SCD at scale. Unlike standard personal information identifiers such as phone numbers, which follow predictable structural patterns, sensitive information like political opinions or health status is inherently contextual in addition to being broad in its definition and reach. Its scope extends beyond explicit statements of sensitive characteristics to information that may reveal those characteristics indirectly. For example, references to a health condition or political belief may appear in a personal disclosure, but equally in reporting, academic discussion or quoted material. The presence of such references alone does not establish that the passage reveals sensitive information about an identifiable person. Attempts to filter sensitive information en masse therefore often lead to high rates of false positives and the exclusion of information necessary for AI model training. Overfiltering personal data or SCD would also have a highly detrimental impact on model quality and safety, including the fact that it would increase the risk of bias and inaccuracies in the model's output.[1]

To accurately identify and filter such data out before collection (i.e., filtering before retrieval), a system would need to read and deeply understand the nuance of every webpage prior to deciding whether to download it. Attempting this in real time across an entire dataset of publicly available data is computationally unworkable. Source-level controls can reduce the likelihood of collecting SCD, but cannot reliably determine whether individual passages reveal sensitive information about an identifiable person. Similarly, mandating the immediate and exhaustive deletion or replacement of all SCD post-collection (i.e., filtering after retrieval but before retention or training) is equally impracticable at scale and ignores the realities of AI data processing pipelines. Furthermore, once collected, text is subsequently broken down into smaller fragments, reorganized, and distributed across vast computing systems. Because finding SCD requires understanding the surrounding context, locating and instantly extracting these specific fragments is operationally unworkable.

Beyond these technical and operational barriers, such requirements could also present heightened privacy risks for users. In practice, large-scale training pipelines ingest unstructured text to learn general linguistic patterns, without seeking to identify specific individuals. Mandating granular pre-collection filtering or immediate post-collection extraction of SCD may require controllers to subject unstructured web data to intensive, identity-level scrutiny. Placing data under this level of individualised inspection risks actively identifying individuals and linking them to sensitive attributes in circumstances where they would otherwise never have been identified. This would run directly counter to the principle of data minimisation and the spirit of Article 11 GDPR. More fundamentally, the EDPB's recognition of incidental and residual processing must retain practical effect. If reliance on the GC and Others framework were

[1] S Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*. See: https://arxiv.org/abs/2305.13169

conditional on measures that effectively require the complete exclusion of SCD, it would provide little accommodation for the very residual processing it recognises. The continued presence of some SCD in a training dataset should not, in itself, establish that a controller's safeguards are inadequate or that the processing falls outside that framework. Rather, the assessment should consider whether the controller satisfies the applicable threshold criteria and has implemented proportionate lifecycle measures (within its responsibilities, powers, and capabilities) to limit incidental collection and mitigate the risks of residual SCD.

We encourage the EDPB to adopt a more flexible approach that accounts for these technical constraints and the associated privacy risks to users. Rather than imposing absolute obligations to pre-filter or instantly replace or delete data or websites as a strict precondition for relying on GC and Others, the Guidelines should require organisations to adopt reasonable and proportionate measures to limit potential processing of SCD.

2. Transparency

We agree with the EDPB's acknowledgement that compliance with the principle of transparency is challenging in the context of scraping large amounts of data from the internet (paragraph 24). We further agree that it is often difficult, impracticable, or objectively impossible to identify and inform data subjects in an effective manner. Consequently, we welcome the EDPB's recognition that the Article 14(5)(b) exemption can be relied upon by controllers facing these logistical barriers.

However, we have reservations about the EDPB's interpretation of Article 14(5)(b) - in particular, the EDPB's suggestion that this exemption should not be relied upon "routinely" by controllers. While we appreciate the EDPB's objective to ensure transparency obligations are robustly upheld, narrowing the availability of the Article 14(5)(b) exemption diverges from the practical framework established by the GDPR. Recital 62 clarifies that assessing "disproportionate effort" is fundamentally a factual and operational exercise, instructing controllers to consider the number of data subjects, the age of the data, and the appropriate safeguards adopted. The phrasing "*in particular*" within Article 14(5)(b) highlights common scenarios where this high threshold is met, but it does not restrict the exemption from being applied to other purposes.

In the context of building datasets for generative AI, the practical challenge of identifying and individually notifying millions of data subjects from unstructured web scraping is immense. Attempting to overcome this challenge would often require developers to scrape additional personal data or cross-reference external databases simply to discover contact details, an exercise that runs directly counter to the principle of data minimisation. This operational reality is identical whether the processing is conducted for the purposes of training generative AI models or archiving related purposes.

We encourage the EDPB to clarify that controllers may rely on this exemption whenever the factual threshold of disproportionate effort is satisfied, removing the restrictive presumption that it cannot be used routinely.

We also have reservations regarding the level of disclosures that the EDPB appears to suggest is required under Article 14(5)(b). In particular, the EDPB's statement that a controller's Privacy Policy should provide a "complete list" of the sources used (or include them to the greatest extent possible) alongside an indication of the "types of sources" that were *excluded* from web scraping and the reasons for their exclusion. Expecting an exhaustive list of included sources, coupled with explanations for excluded sources, is counterproductive for users. Both Art 12(1) GDPR and the Article 29 Working Party Guidelines on Transparency (which have been endorsed by the EDPB)[2] make clear that transparency information must be concise, intelligible, and easily accessible to the average user. Providing pages of granular source lists and technical exclusions is unlikely to be clear or meaningful to everyday individuals and instead risks creating the very information fatigue that the transparency guidelines warn against.

Furthermore, requiring this level of negative disclosure would likely be highly confusing and strays beyond the statutory intention of Art 14 (which focuses on actual processing). For example, if an AI developer provides a summary of the primary categories of public websites it scrapes, but is then forced to append a lengthy annex explaining why it did not scrape other websites or data types, the privacy notice would become dominated by overwhelming and irrelevant information. This would inevitably dilute the core transparency message, making it substantially harder for everyday users to understand how their actual data is being used.

We encourage the EDPB to reconsider these recommendations regarding exhaustive source lists and excluded sources to ensure privacy notices remain concise, intelligible to the average user, accessible, and focused on meaningful disclosures that avoid information fatigue.

We note that the Guidelines also propose, as a matter of good practice, that controllers publicly include domain names and URLs of scraped pages in searchable formats, alongside exact dates and times of collection. While we understand the EDPB's underlying intent to empower data subjects, we are concerned that this proposal for public-facing, granular logging is operationally unworkable for the rapid, large scale data collection required to train modern foundation models. The sheer volume and velocity of ingestion, combined with the continuous, distributed, and dynamic nature of web crawling pipelines, makes capturing,

[2] Article 29 Working Party, Guidelines on transparency under Regulation 2016/679 (paragraph 8).

indexing, and maintaining persistent, publicly accessible, real-time searchable logs of billions of individual URLs practically unworkable.

Publishing granular information about dataset composition also introduces acute security vulnerabilities. Disclosing exact source URLs and collection timestamps provides a detailed roadmap for malicious actors to conduct adversarial attacks, including targeted data poisoning across identified web domains and sophisticated data extraction or inference attacks against downstream models. In addition, details of training data are prototypical examples of proprietary trade secrets (i.e., AI model training recipes); mandating their public disclosure would severely compromise confidentiality by giving competitors direct visibility into proprietary data curation strategies.

Rather than expecting controllers to publish and maintain public-facing, granular searchable logs of domains, URLs, and collection timestamps, the Guidelines should recognise that describing the broad categories of publicly available sources scraped and the web crawlers used provides meaningful, accessible transparency to users without introducing severe operational burdens, security risks, or threats to commercial confidentiality.

We encourage the EDPB to remove the recommendation to publish searchable lists of domain names, URLs, and collection timestamps, and instead adopt a proportionate approach that allows controllers to provide meaningful transparency by describing the broad categories of scraped sources.

We note the EDPB's suggestion that controllers make their Data Protection Impact Assessments (DPIAs) publicly available as a measure to protect the rights, freedoms, and interests of data subjects in the context of large scale web scraping. While we fully support the overarching objective of demonstrating accountability and protecting data subjects, this specific proposal presents significant practical challenges.

By their very nature, DPIAs are detailed internal accountability documents that inherently contain highly confidential and commercially sensitive information. Publishing these documents would necessitate extensive redactions to protect trade secrets and system integrity. As a result, the heavily redacted public versions would likely offer very little meaningful insight and could instead prove disjointed and confusing for readers.

More generally, if there is an expectation on organizations to publish DPIAs concerning large scale web scraping, it would likely result in less rigorous assessments being carried out. If controllers anticipate that their DPIAs will enter the public domain, they will inevitably prepare shorter documents and are far less likely to candidly evaluate residual risks. One of the primary values of a DPIA lies in encouraging open, self-critical risk identification within an organisation.

Expecting public disclosure directly disincentivises that level of candour, which ultimately undermines one of the core purposes of conducting DPIA in the first place.

Dedicated transparency and documentation surfaces, such as an organisation's Privacy Policy, any supplemental privacy notices, as well as public technical reports or model cards, are far more appropriate and effective vehicles for this purpose. These documents are explicitly designed and drafted for external consumption. They translate complex technical and legal processing activities into clear, accessible language, ensuring that individuals can genuinely understand how their data is being used. Relying on these established, user focused transparency mechanisms better serves the GDPR's goal of providing clear information to the public without compromising confidential operational details.

We encourage the EDPB to remove the suggestion that controllers make DPIAs available to the public. Instead, the Guidelines should clarify that established external transparency mechanisms, such as privacy notices and supplemental AI policies, are the appropriate means for informing the public without compromising commercially confidential information or chilling candid internal compliance.

Finally, we welcome the EDPB's express acknowledgement that, in accordance with Article 11 of the GDPR, controllers are not required to maintain, acquire, or process additional information to identify data subjects solely for the purpose of complying with the regulation in the context of web scraping.

3. Legitimate interests

We welcome the EDPB's explicit recognition that controllers can rely on legitimate interests as a valid lawful basis for the collection of data for AI model training. However, we are concerned that the EDPB's commentary on the necessity assessment and the subsequent balancing exercise establishes expectations that are operationally impracticable and do not fully align with the technical realities of foundational model development.

Necessity

The Guidelines suggest that narrowing data collection criteria to target specific information, rather than scraping broad segments of the internet, "*may be crucial*" to satisfying the necessity condition under Article 6 of the GDPR (paragraph 52). However, this approach simply does not align with the technical realities of developing modern generative AI.

Unlike traditional software systems or narrow AI designed for highly specific tasks, foundation models require broad, diverse, and untargeted data ingestion to effectively learn the

complexities of human language, reasoning, and global context. Imposing strict limitations on the scope of data collection would severely degrade model performance, limit dataset diversity, and potentially introduce structural biases by excluding a wide range of human expression.

Moreover, expecting controllers to narrowly target specific information introduces a heightened privacy risk for users. When scraping broad segments of the web to learn general capabilities and linguistic patterns, data is ingested at scale and fragmented without any attempt to inspect individual records or identify specific people. By contrast, requiring controllers to apply targeted collection criteria forces them to closely evaluate and curate specific content, which significantly increases the likelihood of actively identifying individuals in circumstances where they otherwise would have remained unidentified within a broad, untargeted training corpus.

We encourage the EDPB to clarify that the broad ingestion of publicly available web data for training foundational models can satisfy the necessity test under Article 6(1)(f) GDPR.

Furthermore, we encourage the EDPB to more fully recognise the substantial societal, economic, and scientific benefits of generative AI models when assessing the broader legitimate interests balancing test. A restrictive interpretation of the necessity condition that fundamentally conflicts with how these technologies are built risks significantly hindering AI innovation and deployment. We suggest that the balancing exercise should carefully weigh these broader benefits alongside the robust technical and organisational safeguards that developers already deploy to protect data subjects.

We encourage the EDPB to explicitly recognize in the final Guidelines the substantial societal, economic, and scientific benefits of generative AI within the legitimate interests balancing test.

Balancing exercise

Paragraph 66(f) suggests that controllers could create global opt-out lists as a safeguard. While we strongly support providing meaningful controls and respecting opt-out choices, establishing centralized global opt-out lists is technically unworkable and presents significant privacy challenges. To function effectively, such a mechanism would require data subjects to provide extensive information identifying them across various unrelated websites and online platforms. This would force controllers to collect, centralise, and process additional identifying information simply to facilitate the opt-out, which is inconsistent with Article 11 GDPR.

Additionally, the assertion in Paragraph 65 that mitigating measures should not be confused with measures already legally required under the GDPR would benefit from greater nuance. In practice, many technical and organisational measures - such as data minimisation filters, dataset deduplication, and strict access controls - are implemented in specific, enhanced ways during AI development that go well beyond baseline statutory compliance. If a specific technical or organisational measure materially reduces the privacy risks to data subjects, the extent to which it mitigates those risks should be taken into account in the Legitimate Interests Assessment. We encourage the EDPB to allow controllers to holistically account for all risk reducing measures when conducting this assessment.

We encourage the EDPB to reconsider the expectation for centralized global opt-out lists, which would require the collection of extensive additional personal data (and potential identification of individuals) in conflict with Article 11 GDPR.

Paragraph 66(c) proposes that controllers exclude personal data found on freely accessible websites where data subjects cannot be considered "reasonably aware of the accessibility." While we appreciate the EDPB's intention to protect individuals' reasonable expectations, we are concerned that this introduces an individualized, subjective inquiry that is operationally impossible to execute in the context of large scale web crawling.

In practice, assessing what each individual data subject subjectively understands regarding the public accessibility of their information across billions of crawled web pages cannot be achieved at scale. A third-party crawler can only evaluate what is objectively verifiable from a technical standpoint: namely, whether content is publicly available online without authentication barriers or technical access controls.

An individual's subjective awareness of how their information is displayed online is inherently governed by the relationship between the data subject and the primary website publisher, who bears direct transparency obligations under Article 13 of the GDPR. Imputing this subjective evaluation onto a third-party developer who subsequently crawls publicly accessible content is unworkable.

We respectfully encourage the EDPB to reconsider this criterion and clarify that assessments of accessibility should rely on objective, technically verifiable standards rather than individual subjective awareness.

4. Accuracy

The Guidelines impose traditional database expectations onto the training of large scale generative AI systems, which presents significant technical and operational challenges.

Training foundational models requires vast, diverse datasets to effectively capture the nuances of human language and context. In practice, it is impossible to verify whether every piece of personal data contained within such datasets is factually accurate, as there is no single source of truth to compare against at this scale. Furthermore, strictly limiting a training dataset to definitively "accurate and up-to-date" data would severely impact the richness and diversity required to create capable models. Training data naturally and necessarily includes human opinions, historical records, and evolving societal conversations.

Consequently, we have strong reservations regarding the specific measures the Guidelines recommend to ensure data accuracy. The suggestion that controllers should restrict scraping to "reliable sources" such as official registries risks creating heavily biased models that fail to reflect everyday human expression.

Similarly, mandating that controllers timestamp data to track its age and conduct factual verification audits on random samples is technically impractical for rapid, large scale data ingestion. For datasets of this magnitude, random spot checking cannot provide meaningful confidence in the factual accuracy of the overall corpus. Moreover, attempting to factually verify individual data points would frequently require developers to collect additional personal information solely for verification purposes, running counter to data minimisation principles and introducing new privacy risks.

Finally, regarding data minimisation, Paragraph 37 recommends that controllers implement a distinct "data mapping and inventory process." We respectfully suggest that requiring a separate, standalone inventory process risks creating duplicative documentation requirements. The objectives of this mapping exercise are already comprehensively addressed by the requirement to maintain Records of Processing Activities (ROPA) under Article 30 of the GDPR. We encourage the EDPB to clarify that compliance with existing ROPA obligations is sufficient to meet this standard.

We recommend that the EDPB adopt a flexible, context-aware approach to accuracy that reflects the technical realities of training generative models, avoiding prescriptive mandates such as continuous sample auditing, registry-only scraping, or timestamping. Additionally, the EDPB should confirm that compliance with Article 30 ROPA requirements is sufficient to satisfy data mapping expectations.

5. Interplay with EDPB Opinion and the Digital Omnibus

The regulatory landscape governing web scraping and AI is actively evolving, most notably through the ongoing legislative discussion on the EU Digital Omnibus package, which includes significant proposals directly relevant to the matters addressed in the Guidelines, such as the GDPR legal basis for AI development and the regime of SCD processing in AI systems.

As these legislative proposals may lead to the amendment of specific provisions of the GDPR, there is a clear risk of future divergence between the final Guidelines and the post-omnibus legislative framework. If the Guidelines establish strict compliance expectations that are subsequently superseded by the Digital Omnibus, organisations will face significant regulatory uncertainty. Developers risk investing substantial resources into technical and organisational measures that may quickly become redundant or misaligned with the updated statutory requirements.

To mitigate this risk and ensure long term regulatory alignment, we respectfully encourage the EDPB to view these Guidelines as an iterative framework. To ensure regulatory certainty, we request that the EDPB provides organisations with a formal opportunity to make further submissions on any updates to the Guidelines once the omnibus legislative process has formally concluded. This collaborative approach would ensure that the EDPB's guidance remains a durable, legally accurate, and practical resource for industry stakeholders.

We encourage the EDPB to afford stakeholders an opportunity to provide further submissions once the EU Digital Omnibus legislative process concludes, ensuring long-term regulatory certainty and alignment.

The Guidelines expressly state that they build upon and complement Opinion 28/2024 certain data protection aspects related to the processing of personal data in the context of AI models, expecting organizations to read and apply both texts in parallel. On the issue of purpose limitation, for example, the Guidelines refrain from providing substantive guidance and instead refer controllers back to Opinion 28/2024.

From the perspective of regulatory coherence and practical compliance, establishing expectations across two distinct documents, introduces unnecessary complexity for organizations. In particular, it makes it considerably more difficult for organizations to navigate their obligations.

We strongly encourage the EDPB to consolidate its guidance on AI model development and web scraping into a single, comprehensive document.