Let us do IT for you

Poznań, 13.10.2026 r.

Comments by DOIT BIZ Sp. z o.o. on the EDPB public consultations - Guidelines 03/2026 on web scraping in the context of generative AI

Subject Matter Expert Team:

Jolanta Czechowska - Team Lider

Ewelina Felczak

Julia Gil

Approved by:

Tomasz Czarnecki – CEO

Karolina Kuba - Gaździcka – Director of the DPA Department

No.	Paragraphs	Topic
A1	3.2.2 (Transparency), paras. 25–28	Exemption under Article 14(5)(b) GDPR (disproportionate effort)
A2	3.2.2 (Transparency), paras. 30–31	Transparency - list of sources and crawler characteristics
A3	3.2.2 (Data minimisation), para. 36 and Section 3.2.3 (Condition 2), para. 52	Necessity and the scale of the dataset
A4	3.2.3 (Consequences for data subjects), para. 60	"Personal data cannot be easily deleted from a model"
A5	3.2.3 (The role of mitigating measures), para. 66(c)	Different output depending on the user agent
A6	3.2.3 (The role of mitigating measures), para. 66(f)	Opt-out list prior to data collection
A7	3.2.4 "Special categories of personal data", paras. 67–74 (in particular paras. 70–74)	Extension of the GC and Others judgment (C-136/17) to AI scraping and special categories of data

A1 Exemption under Article 14(5)(b) GDPR (disproportionate effort)	
Section / paragraph	Section 3.2.2 (Transparency), paras. 25–28
EDPB position	A controller carrying out web scraping may not be required to inform data subjects individually (Article 14(5)(b) GDPR) where this proves impossible or would involve a disproportionate effort; the balancing exercise is carried out for the dataset as a whole (number of data subjects, age of the data, safeguards adopted). The EDPB states, however (para. 27), that this exception should not be routinely relied upon by controllers who do not process personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes.
Our comment	The caveat in para. 27 in practice excludes most commercial AI scrapers from relying on Article 14(5)(b), since commercial model training does not usually pursue archiving, research or statistical purposes. At the same time, it is in tension with para. 25 and with Example 2, in which the EDPB accepts disproportionate effort precisely for a commercial AI developer scraping a very large dataset (social media, forums, data going back 20 years) - i.e. for a typical commercial rather than research case. This creates a contradiction: the wording of para. 27 suggests that commercial scraping for AI training cannot, as a rule, rely on the exception, whereas Example 2 suggests that it can. Controllers do not know which of these indications to follow.
Justification	Article 14(5)(b) GDPR uses the words "in particular", and Recital 62 GDPR, which refers to the number of data subjects, the age of the data and any appropriate safeguards adopted as assessment criteria, does not confine the exception to the purposes referred to in Article 89(1). Archiving, research

	and statistical purposes are thus examples rather than an exhaustive list of conditions. The strong wording "should not be routinely relied upon" risks narrowing the exception below its statutory wording and effectively making its availability dependent on the purpose of the processing, which the provision does not list as a criterion.
Proposal	We ask the EDPB to clarify the relationship between para. 27 and Example 2. We propose that the EDPB: • confirm that the list of purposes in para. 27 is illustrative (in line with the words "in particular" and Recital 62), and that the balancing exercise for the dataset as a whole (para. 28), rather than the purpose of the processing as such, is decisive; • redraft para. 27, e.g.: "Outside the purposes referred to in Article 89(1), reliance on the exception requires particularly careful documentation of the balancing exercise; this does not, however, mean that the exception is excluded for other purposes." **Question to the EDPB:** • Can commercial scraping for AI training, where informing data subjects individually is objectively impossible (as in Example 2), rely on Article 14(5)(b) GDPR?

A2 Transparency - list of sources and crawler characteristics	
Section / paragraph	Section 3.2.2 (Transparency), paras. 30–31
EDPB position	Where informing data subjects individually proves impossible or would involve a disproportionate effort, the controller must always make information publicly available (e.g. in a privacy policy) covering the elements of Article 14(1)–(2) GDPR, including the categories of data, the purposes and the legal basis, and a precise indication of the source(s), stating whether the sources are publicly accessible and, in the case of crawling/scraping, the crawler's characteristics. Where a complete list of sources is impossible or disproportionate, sources should be listed to the greatest extent possible, together with the types of sources omitted and the reasons for omitting them. As good practice (para. 31), the EDPB recommends providing the domain names and URLs of scraped web pages in a searchable format and the date or period of collection.
Our comment	The requirement of a "precise indication of the source(s)" and of the crawler's characteristics is operationally burdensome and commercially sensitive, and the line between what is mandatory (Article 14) and what is "good practice" (para. 31) remains blurred. • In untargeted scraping (crawling), the controller has limited knowledge of which web pages and data end up in the dataset; drawing up a "precise" list of sources may be objectively unfeasible. • The selection of sources, collection criteria and crawler characteristics usually constitute trade secrets and competitively sensitive information; their publication may reveal the methodology and facilitate the circumvention of website safeguards or selective blocking of the crawler.

	• The publication of complete, searchable lists of domains and URLs may itself create new risks (e.g. facilitate re-identification or aggregation of individuals' data).
Justification	The transparency principle (Article 5(1)(a) and Articles 12–14 GDPR) must be applied with due regard to proportionality and to the protection of trade secrets (by analogy, Recital 63 GDPR, which allows for the protection of trade secrets; Directive (EU) 2016/943 on the protection of trade secrets). In describing untargeted crawling, the Guidelines themselves acknowledge that the controller has "limited knowledge" of the data collected, which is difficult to reconcile with the requirement of a "precise" indication of sources. Moreover, para. 30 frames the "precise indication of the source(s)" and the "crawler's characteristics" as mandatory, whereas para. 31 (domains, URLs, searchable format, dates) is framed as "good practice". This distinction is not sufficiently clear in the text, creating a risk either of excessive disclosure (and loss of trade secrets) or of an allegation of non-compliance with the information obligation.
Proposal	We ask the EDPB to: • clearly separate the mandatory elements (Article 14(1)–(2)) from the recommended good practices (para. 31) and confirm that indicating the categories and types of sources and the general characteristics of the crawler in principle satisfies the information obligation, while a list of specific domains and URLs remains a recommendation; • confirm that the requirement of a "precise indication of the source(s)" is subject to limitation on grounds of trade secrets and security (Recital 63 GDPR; Directive (EU) 2016/943), and describe the admissibility of layered transparency (publicly available information for data subjects, with details made available to the supervisory authority upon request); • specify the minimum scope of the "crawler's characteristics", so as not to require disclosure of information enabling the circumvention of website safeguards. **Question to the EDPB:** • In the case of untargeted scraping, where drawing up a list of domains is objectively unfeasible, is it sufficient to indicate the types and categories of sources together with the reasons for omitting some of them (in line with the last sentence of para. 30)?

A3 Necessity and the scale of the dataset	
Section / paragraph	Section 3.2.2 (Data minimisation), para. 36 and Section 3.2.3 (Condition 2), para. 52
EDPB position	The principle of data minimisation does not prohibit training an AI model with large volumes of data, but requires the controller not to process personal data that would not be necessary, relevant or adequate, and to carry out an assessment before scraping begins (para. 36). Narrowing the collection criteria, rather than scraping a wide part of the internet, may be crucial to ensure that the necessity condition is met (para. 52).
Our comment	Ambiguity as to the boundary between an acceptable dataset scale and excessive data collection. The Guidelines do not provide any method or

	criterion for demonstrating that the scope of scraping adopted was necessary, rather than merely useful for model quality, which leaves controllers with genuine legal uncertainty and hampers accountability.
Justification	Paras. 36 and 52 concern two distinct tests - data minimisation (Article 5(1)(c) GDPR) and necessity under legitimate interest (Article 6(1)(f) GDPR) - which are not clearly linked methodologically. The necessity test requires demonstrating the absence of an "equally effective and less intrusive" way of pursuing the purpose (para. 51(ii)), but the Guidelines do not specify how to distinguish a dataset breadth justified by the specific purpose (e.g. a general-purpose model) from excessive scraping justified solely by a general benefit to model quality. In our audit practice, controllers most often justify the scale of the dataset by model quality metrics rather than by a test of necessity for the purpose.
Proposal	Clarify in the Guidelines the relationship between para. 36 and para. 52, and provide an example methodology or template for documenting the ex ante assessment referred to in para. 37.

A4 "Personal data cannot be easily deleted from a model"	
Section / paragraph	Section 3.2.3 (Consequences for data subjects), para. 60
EDPB position	The EDPB states that in practice it may be difficult for data subjects to exercise their rights and to object to scraping, and that one of the main consequences of web scraping in the context of generative AI is that, given the current state of the art, once the model is trained, personal data cannot be easily deleted from it (para. 60).
Our comment	The statement is a categorical factual assertion about the state of technical knowledge which: • is not supported by a footnote or reference to the source or study on which the EDPB based this assessment, even though it is an important element of the balancing test under Article 6(1)(f) GDPR that weighs against controllers; • does not distinguish between two different technical phenomena: (a) the impossibility of removing the influence of specific data from the parameters (weights) of a trained model (unlearning in the strict sense), and (b) the possibility of limiting the extraction of data at the model output level by means of filters and safeguards - which the Guidelines themselves refer to in para. 74(iii)–(iv); • is worded more categorically than para. 74(iv), which assumes that technological progress may in the future make guaranteed deletion of data from a model possible - pointing to an internal inconsistency between para. 60 and para. 74(iv).
Justification	Current scientific literature largely confirms the accuracy of the EDPB's statement, but with important qualifications. A comprehensive multi-institution paper presented at NeurIPS 2025 ("Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy and Research") indicates that there is currently no method for reliably and selectively removing a specific data item from the weights of a trained model without full

	retraining, and that even retraining from scratch without the data concerned does not guarantee that the information will not remain latent in the model's learned representations. At the same time, this literature consistently distinguishes "removal" (from model parameters) from "suppression" (at output level) - most techniques applied in practice (output filtering, safety fine-tuning, prompt restrictions) address only the latter. This matters legally: the statement in para. 60, worded generally as "cannot be easily deleted from a model", may be read as going further than what the technical literature supports (i.e. the absence of a method for removal from parameters), and may disregard the fact that controllers have at their disposal techniques that reduce the risk of data disclosure at output level, however imperfect. The EDPB itself, in para. 74(iv), recognises the conditional, state-of-the-art-dependent nature of this limitation.
Proposal	• Add to para. 60 a footnote referring to current technical literature on machine unlearning, to make the factual basis of a statement used in the balancing test transparent. • Clarify the wording of para. 60 by distinguishing the removal of data from model parameters from the prevention of its extraction at output level, e.g.: "…given the current state of the art, once the model is trained there is no generally available method for reliably and selectively removing the influence of specific personal data from the model's parameters without retraining it; available techniques (e.g. output filtering, safeguards) may reduce, but do not fully eliminate, the risk of the model reproducing such data". • Align para. 60 with the more open, technology-neutral approach of para. 74(iv), so that the statement does not read as a permanent technological limitation.

A5 Different output depending on the user agent	
Section / paragraph	Section 3.2.3 (The role of mitigating measures), para. 66(c)
EDPB position	As a mitigating measure, the EDPB refers to excluding from scraping personal data on freely accessible sites where data subjects cannot be considered reasonably aware of the accessibility - e.g. where a service returns different outputs depending on the user agent used, and the scraping service receives more data (metadata such as timestamps or access statistics via a technical API) than a regular user.
Our comment	The wording is imprecise and may cover situations not intended by the provision, in particular the lawful, authorised use of official APIs, which by definition return more metadata than the browser interface. Nor is there a clear threshold above which "more data" justifies exclusion.
Justification	Para. 66(c) is an example illustrating a mitigating measure within the assessment of data subjects' reasonable expectations (linked to paras. 62–64), not a technical prohibition. However, the current wording ("the scraping service receives more data … than the regular user") does not distinguish between circumventing safeguards deliberately put in place by the website

	operator (by analogy with robots.txt) and using an officially provided, authorised API channel.
Proposal	Clarify that the measure concerns only unauthorised circumvention of safeguards deliberately introduced by the website, and not the use of officially provided interfaces (APIs).

A6 Opt-out list prior to data collection	
Section / paragraph	Section 3.2.3 (The role of mitigating measures), para. 66(f)
EDPB position	Among the mitigating measures, the EDPB refers to facilitating the exercise of individuals' rights, including a discretionary and prior right to object, and the possibility of creating an opt-out list allowing data subjects to object to the collection of their data on certain websites by providing information identifying them on those websites, including before the data collection occurs.
Our comment	A prior opt-out mechanism requiring self-identification is difficult to implement in practice and creates a vicious circle: using it requires prior knowledge of which controller intends to scrape data, when and from which websites.
Justification	Verifying an opt-out request requires obtaining additional identifying data from the very person who wishes to limit the processing of their data, which runs counter to the logic of para. 29 (no obligation to acquire additional information solely in order to comply with the GDPR) and creates a risk of abuse, e.g. false requests made on behalf of other persons, in the absence of a reliable identity verification mechanism across thousands of unrelated services. For datasets covering thousands or millions of individuals, it is also operationally difficult to implement.
Proposal	Confirm that such a list is one of several possible options rather than a necessary condition, and indicate simpler solutions, e.g. a single, aggregated technical signal (similar to robots.txt) instead of separate requests for each website.

A7 Extension of the GC and Others judgment (C-136/17) to AI scraping and special categories of data	
Section / paragraph	Section 3.2.4 "Special categories of personal data", paras. 67–74 (in particular paras. 70–74)
EDPB position	The EDPB considers that the CJEU judgment in GC and Others (C-136/17), concerning a search engine operator, may be relevant by analogy for the incidental and residual collection of special categories of personal data during web scraping for AI training, provided that a four-condition test (para. 73(a)–(d)) is met, based on the controller acting "within the framework of his responsibilities, powers and capabilities", as specified by four types of measures covering the entire life cycle of the model (para. 74(i)–(iv)).
Our comment	The analogy raises interpretative doubts because:

	• the C-136/17 judgment concerned the application of the prohibition in Article 8 of Directive 95/46 (the counterpart of Articles 9–10 GDPR) solely at the subsequent stage - the verification carried out upon a data subject's request for de-referencing (paras. 47–48 of the judgment) - and not at the stage of initial, large-scale and proactive data collection, which is what web scraping for AI involves; • the four-condition test in para. 73 uses vague notions ("relevant similarities", "only the incidental and residual processing", "difficult/impossible to assess") without operational criteria and - unlike other parts of the Guidelines (cf. Examples 1–7) - without a single example illustrating its application; • the four measures in para. 74(i)–(iv) impose obligations on the controller extending over the entire life cycle of the model (before collection, after collection, during development, after deployment), without an explicit reference to the principle of proportionality in relation to the scale of processing or the controller's resources, which in practice may mean a significant and unquantified compliance burden, especially for smaller entities.
Justification	CJEU judgment of 24 September 2019, C-136/17, GC and Others, ECLI:EU:C:2019:773: in paras. 42–44 the Court expressly rejected the idea of a general exemption of search engine activity from Article 8 of Directive 95/46 (Articles 9–10 GDPR); in paras. 45–46 it held that the specific features of the processing carried out by a search engine operator (no creation or publication of data, merely referencing a third-party web page) may affect the extent of its responsibility; in paras. 47–48 the Court explicitly linked the application of the prohibition to the verification carried out following a request by the data subject - i.e. to a reactive mechanism (the right to erasure, now Article 17 GDPR) and not to the initial indexing or collection of content. This structural difference - processing upon request post factum versus proactive processing ex ante - is legally significant, since the "responsibilities, powers and capabilities" test served the Court to limit the operator's liability at a narrowly defined point in the process (verification following a request), not to establish a general standard of due diligence for the entire processing. The EDPB itself acknowledges the limited scope of the analogy (para. 72), but does not translate this caveat into practical criteria for assessing "relevant similarities" (para. 73(a)).
Proposal	• Develop in paras. 72–73 the reasoning as to why the analogy between reactive processing (verification upon request in C-136/17) and proactive processing (training-related scraping) is justified despite the structural difference between these stages of processing. • Add to Section 3.2.4 at least one example (in line with Examples 1–7), including at least one in which the test (a)–(d) is not met, to improve the predictability of its application. • Clarify condition (a) in para. 73 by indicating criteria for assessing "relevant similarities" (e.g. scale of processing, degree of automation, lack of direct control over the source content, function of the processing in the AI value chain). • Introduce in para. 74 an explicit reference to the principle of proportionality in applying measures (i)–(iv) in relation to the scale of processing and the controller's resources (a risk-based approach), so

	that the requirement is not uniform for large and small entities developing AI.