bitkom

Position Paper

October 2026

Bitkom on the EDPB Guidelines 03/2026 on web scraping in the context of generative AI

Summary

We welcome the EDPB's Guidelines 03/2026 on web scraping in the context of generative AI. The Guidelines provide important clarification on the application of the GDPR to the collection and further processing of personal data obtained from publicly accessible sources for the development and training of generative AI models. We particularly appreciate the EDPB's efforts to address the responsibilities of the different organisations involved in the web-scraping and AI-development chain, as well as the safeguards required throughout the data-processing lifecycle.

At the same time, we would suggest that the Guidelines further clarify certain aspects of the relationship between web scraping, the nature of AI model training, and the GDPR concept of personal data. Additional guidance would be helpful regarding the distinction between the incidental presence of personal data in training material and processing that enables the identification, targeting, profiling, or individual treatment of data subjects. We also suggest that the Guidelines take account of the practical and proportionality-related challenges associated with applying certain obligations to large-scale, automated AI-training processes.

Overarching Considerations

Purpose and nature of AI model training

Web data collected for training AI models is generally processed to identify statistical patterns in language, knowledge, and reasoning. The aim is not to gather information about specific individuals. While personal data may be present in the training material, the resulting model normally consists of abstracted parameters rather than a database

of identifiable persons. Therefore, the Guidelines could distinguish more clearly between incidental personal data content in training material, residual memorisation risks, and processing that enables the identification, targeting, profiling, or treatment of specific individuals.

Contextual assessment of personal data and identifiability

The Guidelines could clarify that the presence of a name, image or other human reference on a webpage does not automatically mean that all scraped content constitutes personal data for the relevant processing operation. In line with Recital 26 of the GDPR, identifiability should be assessed by reference to the means that are reasonably likely to be used by the controller. This should take into account the organisation's technical capabilities, the context in which the data is processed, and its actual ability to single out or act upon an individual. This would enable a more nuanced evaluation of large-scale, automated AI training processes.

Identifiability should also be understood functionally: distinguishability within a dataset is not identification. Recital 26 treats singling out as one example of the means that may be used, not as a test in itself, and identification presupposes that the controller can single out a person and, as a consequence, treat them differently. This relative, controller-relative approach is confirmed in Breyer (C-582/14) and SRB v EDPS (T-557/20).

Proportionality and competitive neutrality

The Guidelines could consider the proportionality and practical effects of the proposed obligations, particularly those relating to transparency. The requirements should provide meaningful protection for individuals while remaining feasible in the context of large-scale AI training. The EDPB could also assess whether obligations imposed on EU-based or EU-regulated developers might put them at a competitive disadvantage compared to providers outside the scope of the Guidelines, without providing an equivalent increase in protection for data subjects. We encourage guidance recognising that there is a range of guardrails and protections that may be effective in protecting privacy interests, depending on the particulars and purpose of crawling and the type of AI training involved, and that therefore flexible frameworks are more appropriate than prescriptive solutions.

Definition of personal data in the context of AI training

The Guidelines explain in paragraph 12 that the GDPR applies to web scraping where the process involves personal data. This may include information that directly identifies an individual, such as a name, photograph, or contact detail, as well as information that may identify an individual indirectly when it can reasonably be linked to them.

Bitkom supports this general starting point. However, the paragraph could provide more clarity on how the concept of personal data should be assessed in the context of large-scale AI training specifically. As it is currently worded, it could be interpreted as suggesting that the mere presence of a name, image or other reference to a person in a training dataset makes the relevant information personal data and could potentially bring the entire scraping and training process within the scope of the GDPR.

That interpretation would not sufficiently reflect the distinction between information that is actually connected to an identifiable living person and information that is processed only as part of general statistical or linguistic material. For example, a common name appearing in running text does not necessarily refer to a specific identifiable individual. Similarly, an image of a person appearing incidentally in a large collection is not necessarily comparable to a structured biometric template or a database designed to identify that person.

Taking a broader view, virtually all web data would effectively be considered personal data by default. This cannot be the correct interpretation of a regulation that establishes rights and obligations based on processing that can affect identifiable individuals. It is not possible to meaningfully provide access to a statistical pattern or rectify a model parameter. Therefore, clarification is needed on the application of the existing GDPR concept rather than on the creation of a separate regime for AI training. The guidelines should clarify that identifiability must be assessed under Article 4(1) and Recital 26 of the GDPR. This assessment should take into account the processing context, the means reasonably likely to be used, and the organisation's capabilities.

For example, the guidelines could confirm that a common first name in running text, an image within a large corpus where the controller has no means of identification and statistical uniqueness within a dataset do not, in themselves, establish that the information relates to an identifiable person.

This assessment should also recognise that information may have been personal data in its original context but may no longer be identifiable for a subsequent organisation using it for a different purpose. In an AI-training pipeline, the organisation may have no identification infrastructure, no intention to identify individuals, and no realistic ability to act upon incidental names or images. The fact that a data point is unique or distinguishable within a dataset should not, on its own, be treated as equivalent to the identification of a living person.

We suggest clarifying paragraph 12 to ensure that the assessment of whether scraped information constitutes personal data is carried out contextually, from the perspective of the controller and in accordance with Article 4(1) and Recital 26 of the GDPR. The mere presence of a human name, image, or other reference in a training dataset should not automatically be regarded as sufficient to establish that the information relates to an identifiable natural person.

The assessment should also consider whether the processing is capable of producing any differential treatment of the individual. Identification that cannot result in any action directed at a person does not engage the protective purpose of the GDPR.

The assessment should take into account the means reasonably likely to be used for identification, based on realistic and foreseeable circumstances, as well as the actual capabilities and processing context of the organisation concerned. In particular, the incidental presence of common names in running text or images of individuals, where the organisation has no realistic means or infrastructure to identify or act upon those individuals, should not automatically render the relevant dataset personal data in its entirety. Statistical uniqueness or the possibility of singling out a data point within a dataset should likewise not automatically be equated with the identification of a natural person.

A similar contextual approach should apply to the designation of special categories of data. Consistent with prior EDPB guidance, it is the context and purpose of the processing, not the mere presence of publicly accessible data capable of revealing a special category characteristic, that determines whether data constitutes a special category of data and triggers the requirement to identify a condition under Article 9(2).

Proportionate Transparency and data subject information requirements

Paragraphs 14, 26-27,30-31,34 and 66(c)

The Guidelines should take greater account of the practical realities of large-scale AI training. Where personal data appears incidentally in publicly available content and the controller has no relationship with the individuals concerned, identifying and contacting millions of data subjects may be impossible or involve disproportionate effort under Article 14(5)(b) of the GDPR. Paragraph 26 should also address a central scenario in this context: third parties named or depicted in content posted by others, whom the controller may be unable to identify or contact using reasonable means. Providing individual notice may first require determining whether, and in which content, the relevant individual's personal data is present in the corpus. This may be particularly difficult where the controller is not required to maintain, acquire or process additional information solely to identify data subjects under Article 11 of the GDPR.

The term »routinely« in paragraph 27 should therefore not discourage reliance on an exception that is available under the GDPR where its conditions are met. The assessment should turn on the concrete circumstances of the processing, including its scale, the nature of the data and the means reasonably available to the controller, rather than on how often the controller relies on the exception**.** Furthermore, the suggestion in Example 2 to exclude data from web pages where data subjects are directly identifiable is unscalable and technologically unfeasible in the context of large-scale web crawling, and should not be treated as a precondition for relying on Article 14(5)(b) of the GDPR.

Transparency requirements should likewise remain proportionate to the scale and nature of the processing. Requiring exhaustive, granular and searchable lists of all scraped URLs, precise dates of collection or the public disclosure of full DPIAs may create significant operational, security and competitive risks , including risks to commercial confidentiality and trade secrets, without necessarily providing

meaningful transparency. Similarly, paragraph 66(c) should not introduce a subjective requirement to determine whether individual data subjects were «reasonably aware» that their data was publicly accessible. Such an assessment cannot realistically be conducted across billions of web pages and millions of individuals. The relevant assessment should instead be based on objectively verifiable features of the relevant source and its accessibility.

The Guidelines should nevertheless clarify that the disproportionate-effort assessment under Article 14(5)(b) of the GDPR remains a case-by-case assessment. In particular, where controllers process data from closed professional communities or domain-specific platforms with an identifiable and limited number of users, the availability of structured means of contacting data subjects should be taken into account. The overall size of an AI training dataset should therefore not, by itself, determine whether the exception applies.

The Guidelines should favour proportionate transparency measures, such as descriptions of domain types or source categories and collection methods, representative examples and appropriate contact channels. The proposed expectation in Example 2 to exclude web pages containing directly identifiable individuals should be deleted. DPIAs should remain confidential and be available to competent supervisory authorities upon request rather than being subject to a general publication requirement. Article 35 of the GDPR requires the assessment but not its publication, and Article 36(3) provides for disclosure to the supervisory authority; a general publication obligation may also undermine the frank identification and documentation of risks. A summary of the key findings and mitigating measures would be more appropriate. The assessment of public accessibility should rely on objectively observable criteria rather than the individual data subject's subjective awareness.

A workable objective criterion would be whether the relevant content was accessible without authentication or other access restrictions through ordinary web browsing at the time of collection. The data subject's awareness is primarily a matter for the website operator when it collects personal data directly from the data subject and must meet its own Article 13 of the GDPR obligations.

Risk-based assessment of impact, accuracy and model architecture

Paragraphs 14,40, 42 and59-60

The assessment of risks and impact should reflect the technical characteristics of AI training and distinguish between the processing of data and actual effects on individuals. In model training, personal data is generally processed to identify statistical patterns rather than to obtain or use detailed information about particular individuals. The mere presence of personal data in a training corpus should therefore not be treated as establishing a concrete impact on the relevant data subject. Paragraph 14 also conflates what is technically possible with what occurs in practice. The risk described depends on memorisation and regurgitation, which developers seek

to detect and mitigate, rather than intentionally enable; the purpose of processing, the model architecture and the safeguards deployed should therefore be relevant to characterising the likelihood and severity of the risk. The balancing assessment under Article 6(1)(f) of the GDPR should focus on the realistic likelihood of concrete adverse effects. In many cases, several independent conditions would need to occur before incidental personal data could affect an individual: the information must be identifiable from the controller's perspective, survive the training process in a relevant form, be memorised by the model, be elicited through a third-party interaction and ultimately reach or affect the data subject. Speculative and low-probability causal chains, particularly where downstream actions are outside the controller's control, should carry less weight than demonstrated or highly probable impacts.

The Guidelines should also distinguish societal-level concerns about mass collection from the individual impact that Article 6(1)(f) of the GDPR assesses in the context of the balancing test (Case C-252/21, paragraph 110). The role of independent third-party users should be considered when assessing causation, foreseeability and the effectiveness of safeguards; it should not, however, exclude the controller's responsibility for risks arising from its own processing. The same risk-based approach should apply to accuracy, both at the model stage and during data ingestion. Requiring factual accuracy of every output containing personal data would not be technically achievable for probabilistic generative AI systems. The obligation should instead focus on appropriate technical and organisational measures, proportionate to the risk, to reduce the generation and dissemination of materially inaccurate personal data. At the data collection and pre-training stage, the application of the accuracy principle should avoid transposing traditional database expectations onto AI model training. In particular, paragraph 42 recommends scraping from »reliable sources«, such as official registries.

However, an interpretation of »reliable sources« that effectively limits training data to official or institutional sources would deprive general-purpose AI models of the breadth of human expression and may result in less diverse and less representative training data. «Reliability» should therefore be assessed in a context-specific manner, taking account of the intended purpose, the type of source, available indicators of quality and the risk associated with the envisaged use. Similarly, suggestions that controllers should apply granular timestamps to scraped data for each item to track freshness and conduct pre-training factual verification audits on random samples should be calibrated to the scale, purpose and technical design of the data-ingestion process. Where scraped personal data is subsequently used for the development or in vitro diagnostic medical devices, the Guidelines should also clarify that the GDPR accuracy assessment operates alongside, and does not replace, applicable product-law requirements concerning data quality, representativeness and validation under Regulations (EU) 2017/745 and (EU) 2017/746.

The Guidelines should also avoid framing the technical difficulty of deleting information from trained models as a compliance failure. Training transforms input data into statistical representations rather than retaining it as a conventional database. Where an individual can potentially be affected, the appropriate remedy should be assessed by reference to the actual point of impact. Output-level safeguards may in appropriate cases provide a more proportionate and technically feasible

response than altering or retraining an entire model. Whether such measures are sufficient should depend on the nature of the risk, the feasibility and effectiveness of available measures, and the rights exercised by the data subject.

The analogy to Google Spain (C-131/12) may be helpful only in a limited sense: the Court required intervention at the point of user access rather than deletion of the source. This supports consideration of measures at the output or dissemination layer, but it does not establish that output-level measures will always be sufficient for AI-model processing. The relevant point of interference may, depending on the facts, be the output or application layer rather than the model parameters themselves. Erasure at dataset level also raises a verification problem: a name in running text may be claimed by anyone bearing it, and verifying an image would require collecting a reference photograph of the requester, in tension with Article 11 of the GDPR.

The Guidelines should clarify that an assessment of impact must consider the actual identifiability of the data, the transformative nature of training, the likelihood of memorisation or regurgitation, the safeguards implemented and the role of independent third-party actions. Similarly, controllers should implement reasonable, risk-proportionate measures to reduce inaccurate outputs, without being required to guarantee the factual accuracy of every model output. At the ingestion stage, expectations for granular timestamping and pre-training factual verification audits should be risk-based and proportionate, and the concept of «reliable sources» in paragraph 42 should be broadened to reflect the necessity of diverse, general internet data to mitigate model bias. The Guidelines should recognise that appropriate safeguards may operate at the output or application layer rather than requiring model-level deletion in every case.

The Guidelines refer to implementing opt-out lists; however, this could be technically challenging and may not deliver the outcome that data subjects expect. It would be helpful for the EDPB to provide further clarity on how it considers that an effective opt-out mechanism could operate in practice. This should include consideration of the complexity for organisations of searching large datasets that may include unstructured and pseudonymised data, the appropriate means of verifying and matching requests, and how data subjects' expectations should be managed transparently, including where complete identification or removal cannot be technically assured.

Where AI training supports regulated medical devices or IVDs, the Guidelines should further clarify that data protection requirements should be assessed in coordination with applicable data quality and validation requirements under the MDR and IVDR.

Legitimate interest and the necessity of broad and diverse training data

Paragraphs 50,52 and 73(a)

The Guidelines appropriately recognise that the development of general-purpose AI models may constitute a legitimate interest even where specific downstream uses have not yet been determined. This should be reflected consistently in the assessment of necessity. Controllers should not be required to identify every future application of a

general-purpose model or conduct separate assessments for each potential use case. A single legitimate interest assessment may cover the general objectives of developing a model that can serve multiple commercial, research, public-interest, internal or external purposes, including models released under open-source or open-weight licences, provided that those purposes are described with sufficient specificity and remain within the scope of the original assessment. A single model may serve commercial, research, public-interest, internal and external objectives simultaneously, and release under an open-source or open-weight licence may itself constitute a legitimate interest.

The reference to narrowing collection criteria in paragraph 52 should likewise not be understood as suggesting that broad web collection is inherently incompatible with the necessity requirement, nor should the Guidelines implicitly favour targeted scraping with restrictive criteria over untargeted internet crawling. Broad, untargeted web crawling may be necessary and proportionate, depending on the model's purpose, architecture and intended capabilities, for training general-purpose and foundational AI models to understand human language and context. The breadth and diversity of training data may be necessary to develop accurate, safe and non-discriminatory AI systems. The assessment should distinguish between the volume and diversity of the overall training corpus and the personal data incidentally contained within it. The relevant question is whether the processing of personal data is necessary for the overall training objective, rather than whether every individual website or data point is individually necessary. This does not remove the need to consider reasonably available and less intrusive measures, including data minimisation measures tailored to the risks of the specific processing. Similarly, prioritizing synthetic data cannot in all cases substitute for real-world data. Synthetic data is itself generated by models that must first be trained on real data, so it cannot wholly replace the diverse real-world corpus that may be required to train the underlying model.

Paragraph 52 also sits uneasily with paragraph 36 and conflates data with personal data: the breadth of a corpus reflects linguistic and cultural diversity rather than necessarily reflecting a need to collect more personal data. It should be read consistently with the AI Act's requirements on representative datasets (Article 10) and non-discrimination objectives (Article 9). The Guidelines should acknowledge that representative and diverse training data can support both model quality and the mitigation of discriminatory outcomes, while ensuring that this does not displace the GDPR's separate necessity and balancing requirements.

We welcome the application of the GC & Others (C-136/17) framework to the incidental collection of special-category data during AI development. Assessing processing within the bounds of a controller's actual »responsibilities, powers, and capabilities« is particularly important for large-scale web scraping.

The Guidelines should also clarify the relevance of the CJEU's reasoning in *GC & Others* to AI training. The relevant similarity with search engines should concern the structural characteristics of the processing, namely, automated, large-scale and potentially undifferentiated collection from publicly available sources, rather than the different downstream purposes of search indexing and AI model training. The existence of these

similarities should be assessed in light of the specific processing operation and should not be presumed solely because data are obtained through web scraping.

For domain-specific AI training, including training for medical language models or clinical decision-support systems, the Guidelines should further clarify that the application of the *GC & Others* reasoning requires an assessment of the specific processing context. The fact that a training activity is domain-specific and directed towards a particular professional or technical purpose should be taken into account when assessing whether the relevant structural similarities with search-engine processing are present. Where special-category data under Article 9 of the GDPR is involved, controllers should be able to demonstrate and document the specific reasoning supporting the application of the relevant legal framework.

The Guidelines should clarify that necessity is assessed in relation to the overall training objective, while taking due account of the role that data diversity in model quality, safety and fairness. For paragraph 73(a), the Guidelines should clarify that »relevant similarities« concern the automated, large-scale and undifferentiated nature of the collection and the inability to fully predetermine the categories of data that will be encountered.

Paragraph 54 – Chilling effect

The chilling-effect concept derives primarily from surveillance contexts in which individuals know they are monitored by an identifiable observer who may act on what is observed. Its relevance to web scraping for AI training should therefore not be assumed. Collection for AI training does not necessarily involve direct observation of, or interaction with, identifiable individuals, and no empirical evidence of self-censorship attributable to AI training is cited. The weight given to this factor should accordingly be based on the concrete circumstances of the processing and supported by evidence, rather than inferred solely from the fact of large-scale collection.

Public accessibility, data mapping and technical opt-out signals

Paragraph 37 and 66

Robots.txt and ai.txt should not be treated as data minimisation requirements under Article 5(1)(c) of the GDPR. These mechanisms primarily reflect website operators' preferences regarding automated access and are more appropriately considered when assessing reasonable expectations and the balancing of interests. In particular, ai.txt does not currently constitute an established or consistently implemented technical standard.

Robots.txt was also designed as crawling etiquette, has no defined legal status under EU law, and is in practice often misconfigured or drafted so broadly that it blocks all automated access. Its presence or absence should therefore be one relevant factual element in the balancing test, rather than a conclusive indicator of either the necessity of processing or data subjects' reasonable expectations.

At the same time, within the context of the balancing test, relying on existing, standardised machine-readable technical opt-out signals made available by website operators (such as robots.txt) is far more workable than the suggestion in paragraph 66(f) to establish centralised, global opt-out lists allowing individuals to object prior to collection. Creating and administering global pre-collection opt-out lists for unstructured web data would place an impossible identification and administrative burden on controllers, effectively requiring them to identify individuals across websites in advance. Such lists would also raise substantial questions regarding the reliable verification, matching and secure handling of the identifiers necessary to give effect to an objection. The Guidelines should clarify that the information maintained in an Article 30 ROPA may satisfy this expectation where it contains sufficient information on the relevant scraping operation, categories of data, sources, purposes, recipients, retention and safeguards. Controllers should not be expected to maintain a separate, duplicative inventory solely because data are obtained by web scraping.

Therefore, the discussion of robots.txt and ai.txt should be moved to the balancing test section, focusing the data minimisation analysis on the necessity and adequacy of the data for the stated processing purpose. In addition, paragraph 66(f) should not be read as requiring centralised, global opt-out lists. Instead, the Guidelines should promote technically feasible mechanisms, including established and machine-readable signals provided by website operators, while recognising their limitations and the need for context-specific assessment. The Guidelines should also explicitly clarify that maintaining an existing ROPA under Article 30 of the GDPR can satisfy the data mapping and inventory expectations of paragraph 37 where it provides the necessary operationally relevant information.

Relatedly, the EDPB should consider providing clarification on the implications of reliance on legitimate interests as a legal basis, in a situation where a controller collects personal data beyond the scope reasonably anticipated in its legitimate-interest assessment—for example, through access to a logged-in environment or despite technical measures indicating that automated collection is not permitted. The Guidelines should clarify that such circumstances require prompt reassessment of the legal basis, necessity, balancing test and applicable safeguards, as well as appropriate remediation proportionate to the circumstances. Proportionate and technically feasible safeguards

Paragraphs 74(ii)–(iv) and 77

While the application of the CJEU's GC & Others framework to incidentally collected special-category data (SCD) is a constructive starting point, the specific conditions outlined in paragraph 74 remain overly rigid and insufficiently aligned with current technical capabiities. In particular, paragraph 74(ii) demands that incidentally collected SCD be »deleted from datasets immediately«. Detecting and removing all special-category data at the point of collection from open-web sources cannot be guaranteed with current technologies, particularly where the data are unstructured, contextual or ambiguous. Crucially, the Guidelines risk creating an unworkable double compliance burden by demanding complete exclusion of SCD at the point of ingestion while simultaneously requiring effective safeguards at the deployment stage to prevent the model from outputting SCD. The assessment should instead consider the combined

effectiveness of safeguards throughout the AI lifecycle. The Guidelines should also distinguish between obligations that can realistically be fulfilled by model providers and those that concern deployers of AI systems. The requirement to »constantly monitor« AI outputs in paragraph 74(iv) is undefined and may be impossible for providers of open-source models or models deployed by independent third parties. Even for proprietary systems, continuous monitoring of all outputs at scale may itself create significant technical and privacy concerns. Monitoring obligations should therefore correspond to the provider's actual role, control over the system and ability to implement effective measures.

The AI Act already distinguishes providers of general-purpose AI models from deployers, and Article 53 sets out specific obligations for providers of general-purpose AI models; deployment-stage monitoring should not fall on providers who do not operate the system. The Guidelines should therefore avoid assigning operational obligations to an actor that lacks practical control over the deployment environment.

Similarly, the obligation to »constantly monitor« technological developments should be framed as a proportionate and periodic review obligation rather than an open-ended requirement to adopt every emerging technique. In particular, promising research should not automatically be treated as the technological state of the art where it has not been reliably validated at production scale.

The EDPB should introduce the concept of »reasonable steps« into paragraph 74, shifting the standard in paragraph 74(ii) from an absolute guarantee of immediate deletion to a workable risk-mitigation framework. The Guidelines should recognise that a holistic combination of reasonable ingestion-phase safeguards and deployment-phase output filtering may, depending on the circumstances, constitute sufficient mitigation for incidentally collected special-category data, rather than requiring technological perfection at both stages independently. Furthermore, the term »constantly monitor« should be replaced with an obligation to conduct reasonable, risk-proportionate monitoring and periodic reviews of relevant technological developments. The Guidelines should distinguish between the responsibilities of model providers and deployers and recognise that techniques should only constitute the »state of the art« where they have been sufficiently validated and are reasonably practicable in the relevant context. A technique should not be considered state of the art solely because it has been described in research; it should also have demonstrated reliability, maturity and practicability for the relevant deployment context. Periodic review intervals should be risk-based; an annual review may be appropriate as a minimum baseline for many processing operations, supplemented by reassessment following material changes in risk, technology or deployment. In paragraph 77, the reference to unlearning algorithms seems premature considering the current state of the art.

The EDPB should consider whether this sets unrealistic data subject expectations in relation to the right to erasure. Currently, robust output filters and other application-level measures may, in appropriate circumstances, be used to give effect to erasure requests, while unlearning algorithms, although the subject of active research and development, have not yet demonstrated consistent technical and commercial feasibility across model types and scales. The Guidelines should clarify that the

appropriate measure must be determined case by case, taking account of the source of the risk, the model architecture, the effectiveness of available mitigations and the specific right exercised.

Example 7- a realistic application of the balancing test

Example 7 should provide a realistic illustration of how a responsible AI developer can satisfy the balancing test rather than setting cumulative conditions that are unlikely to be achievable in practice.

In particular, the reference to copyright compliance should be removed. Copyright and data protection are distinct legal regimes with separate legal bases, authorities and enforcement mechanisms. Compliance with copyright law may be relevant to a controller's activities generally, but should not be presented as a condition for establishing a legitimate interest under Article 6(1)(f) of the GDPR. The EDPB's mandate is the consistent application of the GDPR. Presenting compliance with Directive (EU) 2019/790 as a necessary element of the Article 6(1)(f) assessment would risk extending the balancing test beyond its GDPR basis. The reference to data subjects having »manifestly made the content public« should likewise be reconsidered. This wording appears to import the specific threshold of Article 9(2)(e) of the GDPR into the Article 6(1)(f) balancing test. For legitimate interest assessments, public accessibility is relevant to the reasonable expectations of data subjects, but it should not be treated as requiring that the data subject personally published the content. The controller should assess, in the circumstances of the individual source, the extent to which data subjects could reasonably expect collection and reuse of publicly accessible content for AI training.

The suggested link between general media coverage of AI training and data subjects' reasonable expectations should be expressed with caution. Broad public awareness may be a contextual factor but cannot by itself establish an individual's reasonable expectation of specific processing. Within the Article 9 analysis, the Guidelines should not assume that all personal data on an open website have been »manifestly made public« by the data subject. Article 9(2)(e) requires an assessment of whether the individual data subject deliberately made the specific special-category data public; public accessibility alone will not necessarily meet that threshold.

More generally, the example should demonstrate that the balancing test can be satisfied through a combination of proportionate safeguards, rather than requiring an idealised model using only public-domain content, exclusively self-published information, comprehensive URL-level transparency and extensive controls that may not be technically achievable at scale. Example 7 should be presented as an illustrative pathway, rather than as a cumulative checklist or exhaustive model of compliance.

Rather than stating that no production model meets all conditions, Example 7 would be more useful as one of several pathways to satisfying the balancing test.

A proposed change should replace the current example with a scenario in which the controller uses publicly accessible, licensed and synthetic data sources, applies measures such as deduplication and other safeguards against memorisation,

implements output-level safeguards, facilitates data subject rights where appropriate and feasible, and provides meaningful, category-level transparency. The example should make clear that safeguards must be selected, documented and periodically reviewed in light of the particular model, processing purpose, data categories and residual risks. This would demonstrate that a genuine risk-based compliance framework with layered safeguards can satisfy the balancing test without imposing requirements that are unrelated to Article 6(1)(f) of the GDPR, technically unrealistic, or disproportionate to the actual risk.

Relationship with the AI Act and other applicable regulatory frameworks

The Guidelines should expressly clarify their relationship with Regulation (EU) 2024/1689 (AI Act). The Guidelines address the lawfulness of web scraping under the GDPR and should not be understood as replacing or determining compliance with separate obligations applicable to AI systems and AI models. In particular, the Guidelines should acknowledge the relevance of the AI Act requirements concerning training, validation and testing datasets, including Article 10 and, where applicable, the obligations applicable to providers of general-purpose AI models under Article 53.

Where the same processing activity is subject to both instruments, controllers and providers should assess the applicable requirements cumulatively, while recognising that compliance with one legal framework does not in itself establish compliance with the other. The Guidelines should also clarify that the classification of a particular AI system under the AI Act does not determine the lawfulness of the underlying processing under the GDPR.

Relationship with the European Health Data Space

The Guidelines should clarify their relationship with Regulation (EU) 2025/327 on the European Health Data Space (EHDS). Where web scraping involves health data within the meaning of Article 4(15) of the GDPR, the Guidelines should acknowledge that certain data may also fall within the material scope of the EHDS, provided that they qualify as electronic health data and the relevant EHDS conditions are met. The Guidelines should therefore recognise that GDPR obligations may operate alongside EHDS-specific rules governing access to, and secondary use of, electronic health data. In particular, Article 53(1)(e)(ii) of the EHDS expressly refers to the secondary use of electronic health data for the training, testing and evaluation of algorithms, including in medical devices, in vitro diagnostic medical devices, AI systems and digital health applications. The Guidelines should make clear that this provision demonstrates the relevance of the EHDS framework for certain AI-development activities involving electronic health data, without predetermining whether a particular web-scraping operation falls within that framework. Transparency expectations under the Guidelines should be interpreted consistently with applicable transparency and documentation

obligations under other EU instruments, including the AI Act where relevant, so as to avoid duplicative or conflicting disclosure obligations.

The Guidelines should therefore make clear that a GDPR-compliant web-scraping practice does not, by itself, determine compliance with applicable EHDS access and use requirements. Conversely, compliance with EHDS-specific access, authorisation or secondary-use conditions does not remove the need to comply with the GDPR.

Proposed change: Include an explicit »without prejudice« clarification stating that the Guidelines do not address or determine the applicability of EHDS-specific access, authorisation and secondary-use mechanisms. The clarification should also state that the applicability of EHDS requirements depends on the data category, the relevant actor and the specific processing context.

Clarification regarding data shared within a corporate group

The Guidelines should clarify the scope of the reference to »own data« in paragraph 5. Where personal data was originally obtained through external web scraping and is subsequently shared between separate legal entities within a corporate group, the data should not become «own data» of the receiving entity merely as a result of the intra-group transfer. The receiving entity should remain subject to the applicable requirements governing the processing of scraped personal data, unless the data has been lawfully anonymised.

This clarification would be particularly relevant for multinational groups with multiple legally independent entities, where data may be shared between business units or subsidiaries for model development or validation. It would help prevent uncertainty as to whether an intra-group transfer changes the legal character of the data while preserving the distinction between personal data and genuinely anonymised information.

Future review in light of legislative developments

The Guidelines should contain an explicit review clause acknowledging that the legal framework for AI and data protection is developing rapidly. In particular, if the EU legislator establishes a specific legal basis or otherwise adopts specific rules governing the processing of personal data for the development or technical operation of AI systems and models, the EDPB should review and, where necessary, update these Guidelines.

This is particularly important in light of ongoing legislative developments, including the Digital Omnibus. The review should cover, in particular, the Guidelines' analysis of legitimate interest, special-category data, transparency and data subject rights. A review mechanism would provide greater legal certainty for organisations that must establish long-term data governance and compliance frameworks for AI development.

Bitkom represents more than 2,200 companies from the digital economy. They generate an annual turnover of 200 billion euros in Germany and employ more than 2 million people. Among the members are 1,000 small and medium-sized businesses, over 700 start-ups and almost all global players. These companies provide services in software, IT, telecommunications or the internet, produce hardware and consumer electronics, work in digital media, create content, operate platforms or are in other ways affiliated with the digital economy. 82 percent of the members' headquarters are in Germany, 8 percent in the rest of the EU and 7 percent in the US. 3 percent are from other regions of the world. Bitkom promotes and drives the digital transformation of the German economy and advocates for citizens to participate in and benefit from digitalisation. At the heart of Bitkom's concerns are ensuring a strong European digital policy and a fully integrated digital single market, as well as making Germany a key driver of digital change in Europe and the world.

Published by

Bitkom e.V.

Albrechtstr. 10 | 10117 Berlin

Contact person

Elena Kouremenou | Manager for Sector Specific Data Protection & Governance

P +49 30 27576-425 | e.kouremenou@bitkom.org

Responsible Bitkom committee

WG Data Protection

Copyright

Bitkom 2026

This publication is intended to provide general, non-binding information. The contents reflect the view within Bitkom at the time of publication. Although the information has been prepared with the utmost care, no claims can be made as to its factual accuracy, completeness and/or currency; in particular, this publication cannot take the specific circumstances of individual cases into account. Utilising this information is therefore sole responsibility of the reader. Any liability is excluded. All rights, including the reproduction of extracts, are held by Bitkom.