Comments of ANGIF, Associazione Nazionale Giuristi Informatici e Forensi on the *Guidelines 03/2026 on web scraping in the context of generative AI* of the European Data Protection Board (version 1.0, adopted on 7 July 2026 for public consultation)

CONTENTS: 1. Introduction: About ANGIF; 2. Scope, chain of actors and territorial scope; 3. Taxonomy of operations and structure of the pipeline; 4. Purpose, public accessibility and reasonable expectations; 5. Technical signals of objection and opt-out mechanisms; 6. Transparency and Article 14 GDPR; 7. Data minimisation, synthetic data, anonymisation and pseudonymisation; 8. Accuracy, reliability of sources and necessity; 9. Legitimate interest, general-purpose models and structure of the balancing test; 10. Data subject rights after training; 11. Data protection impact assessment, FRIA, data protection by design and security; 12. Coordination with the Artificial Intelligence Act and copyright law; 13. Special categories of personal data and the GC and Others judgment; 14. Output monitoring, state of the art and documentation of the supply chain; 15. Terminology, protection of minors and editorial remarks; 16. Concluding remarks.

Abstract

ANGIF's comments focus on the gap between the principles stated in the Guidelines and their application in the examples, and on the verifiability of compliance. In particular, ANGIF asks the EDPB to: introduce an explicit duty of due diligence on acquired datasets (Par. 21); reword Examples 2, 3, 4 and 7, which concede more than the corresponding principles allow; state that the absence of a signal of objection can never indicate reasonable expectation (Par. 64); recognise the DPIA under Article 35 GDPR as a rule mandatory, given the limited scope of the FRIA under Article 27 of the AI Act; clarify when trained models are anonymous and how data subject rights apply to them (Par. 60); address Articles 9(2)(e) and 10 GDPR; coordinate the Guidelines with Directive (EU) 2019/790 and the AI Act; and develop a standardised, forensically verifiable dataset provenance record.

1. Introduction: About ANGIF

ANGIF, the Associazione Nazionale Giuristi Informatici e Forensi (Italian National Association of Legal Informatics Jurists and Digital Forensic Examiners), brings together jurists, lawyers, university lecturers, a PhD holder in legislative drafting and legimatics, and digital forensic examiners working in information technology law, legal informatics, personal data protection and digital forensics. Its statutory purposes include the study of the legislation and legal methodologies relating to information and communication technologies, the techniques for the detection and prosecution of offences committed by means of information technology, the training of legal practitioners and dialogue with national, European and international institutions. Many of its members act as technical experts in judicial proceedings: the comments that follow therefore reflect, in particular, the perspective of those who have to verify in court, and not only declare, what a system has actually done.

Since the applicable obligations follow, in most cases, with sufficient clarity from the GDPR, the added value of the Guidelines lies in their ability to translate principles into operational criteria that a controller can apply *ex ante* and a supervisory authority can verify *ex post*. The comments below are placed on this level. Particular attention is paid to the examples, because practitioners read the examples of an interpretative document as precedents: an example that concedes more than the principle allows risks becoming, *de facto*, the criterion actually applied. Each remark refers to the paragraph or example concerned, with the numbering of version 1.0.

2. Scope, chain of actors and territorial scope

Due diligence on acquired datasets. Par. 3-6 exclude from the scope of the Guidelines data brokers who scrape data without training models themselves (Par. 4), while including organisations that obtain and reuse a dataset already scraped by a data broker (Par. 3, second scenario); Par. 19 adds that the scraper is not, in principle, responsible for the re-use of the data. The burden of verifying the lawfulness of the original collection therefore falls entirely on the acquiring developer, which, under Par. 21, is merely required to "consider" the datasets it intends to use. ANGIF considers that this verb

is too weak and proposes that the Guidelines set out an explicit duty of due diligence, under which the re-user may not rely on any *praesumptio* of lawfulness of the original collection, not even a rebuttable one, but must carry out a positive assessment, proportionate to the circumstances, of: the provenance of the data; the period and methods of collection; the categories of sources and the selection criteria; the legal bases declared; the minimisation and filtering measures applied; compliance with signals of objection; the chain of intermediate transfers. EDPB Opinion 28/2024 of 17 December 2024 already addresses the consequences of unlawful processing in the development phase for subsequent controllers; the reference to it in Par. 23, currently limited to purpose limitation, should be extended to this aspect.

Dataset licensors. Par. 16-20 and Example 1 cover the scraper acting as processor (Par. 18), separate processing activities (Par. 19) and joint controllership (Par. 20). They do not cover what is probably the most common case in practice: a provider that builds a dataset according to its own criteria and licenses it to several developers, sometimes tailoring it to each of them. Here the boundary between independent controllership, joint controllership and processing on behalf of the controller is thin, and a dedicated example is suggested.

Territorial scope. The Guidelines do not address the territorial scope of the GDPR, although a significant part of scraping is carried out by entities established outside the Union. It should be clarified whether, and under what conditions, the automated collection of data relating to data subjects in the Union falls within Article 3(2)(b) GDPR as monitoring of behaviour, with the consequent obligation to designate a representative under Article 27. It should also be stated that, in line with Guidelines 05/2021 on the interplay between the application of Article 3 and the provisions on international transfers as per Chapter V of the GDPR (version 2.0, adopted on 14 February 2023), direct collection by a non-EU entity is not in itself a transfer, whereas flows between a scraper established in the Union and a developer established in a third country, and *vice versa*, may be; the public accessibility of the source is irrelevant to both questions.

3. Taxonomy of operations and structure of the pipeline

The sequence in Par. 11 (definition of collection criteria, extraction, cleaning, structuring and storing) stops at the creation of the dataset. Yet the assessment of many issues addressed later in the document (legal basis, storage, exercise of rights, memorisation) depends on the subsequent phases: pre-training, fine-tuning and post-training, evaluation of the model and retrieval of content at inference time. It is suggested that Par. 11 be supplemented by a functional description of these phases, to which Par. 14, 60 and 74 already implicitly refer, so that rules which operate differently along the chain are not attributed to a single operation called “scraping”.

Two technical distinctions drawn by the document should also be carried through to the balancing test. First, Par. 8 recognises that untargeted scraping increases the risk that the controller has limited knowledge of what personal data are collected, which makes it harder to meet the accountability obligation under Article 5(2) recalled in Par. 14; Par. 57 should draw the consequence that indiscriminate collection in itself weighs against the controller, unless documented compensatory measures are in place. Second, Par. 9 describes crawlers of dynamic content as mimicking a human user’s interaction with the page, and Par. 66(c) regards as critical the case in which the scraper receives more data than the regular user. Read together, these passages show that simulating human behaviour is also the means of circumventing the signals of objection that Par. 37 and 64 rely on; the connection should be made explicit.

4. Purpose, public accessibility and reasonable expectations

Purpose limitation. Par. 23 deals with purpose limitation in a single sentence, referring to Opinion 28/2024. In this context, however, the gap between the purpose of publication and the purpose of collection is structural: information disseminated for professional, informative, social or personal reasons is incorporated into a corpus for a GP model, from which further uses arise that cannot be foreseen at the time of collection. The Guidelines should distinguish between the purpose of the original publication, that of the scraping, that of the training and those of the subsequent uses of the

model, and state that the mere availability of data online does not allow the compatibility of further processing to be inferred, either *expressis verbis* or *per facta concludentia*.

Public accessibility. The notion of publicly accessible data is used (Par. 64, 66(c), Example 7) as if it had clear-cut boundaries. Content may be public and indexed, public but reachable only through its address, available after free registration, accessible through APIs, stored in caches or archives, or technically retrievable by a crawler but practically impossible for a regular user to find. Technical accessibility, practical accessibility, the accessibility intended by the publisher and that reasonably perceived by the data subject are different things, and the balancing test should keep them apart.

Reasonable expectations. Par. 64 states, as a "rule of thumb", that data made public by the data subject themselves are more likely to be reasonably expected to be processed by others. The statement holds *in abstracto*, but must be read *cum grano salis*: a curriculum vitae, a family photograph, a political statement, a scientific article, an account of a health condition and a forum comment are published in unrelated contexts and generate entirely different expectations as to their use for training a generative model. The criterion of contextual integrity, which assesses an information flow against the context in which the information was shared, would give Par. 64 a more reliable parameter than the mere public character of the data. Par. 64 should also distinguish between data made public by the data subject, data published by third parties with or without the data subject's consent, and data published by the data subject in one context and replicated by others elsewhere. The reference to public figures should specify that public notoriety affects the balancing test in the specific case, without generally lowering protection.

Example 4. Example 4 conflicts with Par. 63. Par. 63, referring to paragraph 53 of Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, states that the mere fulfilment of the information obligations under Articles 12, 13 and 14 GDPR is not sufficient in itself to consider that data subjects can reasonably expect a specific processing. Example 4 nevertheless concludes that users can reasonably expect third parties to scrape their data for AI development merely because the platform has communicated that possibility to them, typically in its terms of service. If the principle in Par. 63 applies to the controller's own information, it applies *a fortiori* to information provided by a third party, the platform operator, which has no relationship with the scraper. Example 4 should be reworded so that the platform's information is one element of the balancing test, not a sufficient one.

5. Technical signals of objection and opt-out mechanisms

Distinguishing the instruments. The Guidelines rely on robots.txt, ai.txt, authentication and CAPTCHA in three places: as a criterion for excluding sources (Par. 11 and 37), as an element reducing reasonable expectations (Par. 64 and Example 5) and, correctly, to exclude that their absence amounts to consent (Par. 45). The four instruments differ in nature and function and should be distinguished: robots.txt is a standardised protocol (IETF, RFC 9309, Robots Exclusion Protocol, September 2022) by which the website operator expresses preferences on automated access; ai.txt is not, at present, an established standard, and the Guidelines should say so rather than place it on the same footing; CAPTCHA often serves security or traffic management purposes unrelated to AI training; authentication may protect very different interests.

Attribution of the signal. The website operator, the data subject, the author of the content, the copyright holder and the person who published the content often do not coincide, and a preference expressed by the operator cannot be attributed *sic et simpliciter* to the data subject. Par. 64 implicitly recognises this, since it makes the signal relevant only where data subjects are aware of the measures adopted by the website. ANGIF submits that the resulting lopsidedness should be stated expressly: the presence of a signal of objection must operate as a minimum threshold and entail the exclusion of the source, as Par. 37 already provides, whereas its absence can never be read, even indirectly, as an indication of reasonable expectation. The General-Purpose AI Code of Practice published by the European Commission on 10 July 2025 points in the same direction, since its copyright chapter commits signatories to employing crawlers that respect robots.txt.*Opt-out mechanisms.* Par. 66(f) provides for a discretionary and prior right to object and for an opt-out list available even before

collection. It is proposed that the Guidelines promote interoperable, machine-readable mechanisms allowing data subjects to express differentiated preferences regarding training, fine-tuning, use in retrieval systems and the generation of outputs relating to them. Such mechanisms must remain distinct from the reservation of rights under Article 4(3) of Directive (EU) 2019/790, which must be expressed "in an appropriate manner, such as machine-readable means in the case of content made publicly available online": the *ratio*, the legal basis and the entitled party are different, and the two are frequently confused in practice.

Processing of opt-out data. The Guidelines do not address the fact that the opt-out system itself entails processing of personal data, since excluding someone requires knowing who they are (names, e-mail addresses, user names, URLs or other identifiers). This processing must be designed in accordance with data minimisation and data protection by design, with a strict purpose limitation preventing the reuse of the identifiers for any purpose other than exclusion.

6. Transparency and Article 14 GDPR

Example 2. Par. 24-34 allow recourse to the exception in Article 14(5)(b) GDPR where individual information is impossible or would involve disproportionate effort. *Ad impossibilia nemo tenetur*, and the principle is not in question; its application in the examples is. Par. 27, in line with WP260 rev.01, states that the exception should not be routinely relied upon by controllers who are not processing personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes. Example 2 nevertheless grants it to a developer training a model with no reference to those purposes, solely on the basis of the size of the dataset and the age of the data. As worded, the example turns the exception into the rule for any large-scale scraping, that is, for the very case the Guidelines are meant to regulate. Example 2 also rests on the premise that no direct identifiers are included in the data, which is hard to reconcile with Par. 12 (the GDPR applies to data that indirectly identifies an individual) and with Par. 54 (the risk that the model is used to identify data subjects who intended to express themselves anonymously).

Parameters of disproportionate effort. The parameters in Par. 28 (number of data subjects, age of the data, safeguards adopted, drawn from Recital 62) should be supplemented with further verifiable elements: the actual degree of identifiability of the data subjects, the availability of contact channels, the granularity of the collection, the level of risk of the processing and the existence of alternative measures reducing the informational detriment. The more the assessment relies on documented elements, the less the exception lends itself to being used as a matter of convenience.

Example 3. Example 3 correctly rules out disproportionate effort for a closed user group of 5 000 individuals identifiable through full names, email addresses and unique participant IDs, but takes the collection itself for granted. Par. 66(c), however, identifies precisely data accessible only when logged in as candidates for exclusion. Without clarification, the example may be read as meaning that scraping a closed group is lawful provided individual information follows, a conclusion which ANGIF does not believe the EDPB intends to endorse.

Article 11 GDPR. Par. 29 recalls that, under Article 11 GDPR, the controller is not obliged to acquire additional information to identify the data subject and that, in that case, Articles 15 to 20 do not apply unless the data subject provides such information. In the scraping context this risks becoming a shield: a controller declaring itself unable to identify the data subjects in the dataset may nonetheless have a model capable of returning identifying information on request. It should be specified that Article 11 cannot be invoked where identification, although not pursued, is made possible by the capabilities of the trained model.

Structured notice. Par. 30-31, on publishing the sources, the crawler's characteristics, domain names and URLs in a searchable format and the period of collection, offer a basis for standardisation. ANGIF proposes that the Guidelines encourage a structured, machine-readable notice in an open format, indicating at least: the identity of the controller; the crawler and its user agent; the purposes; the categories of data; the legal basis; the period of collection; the categories of sources; the storage period; the contact point for the exercise of rights. Article 53(1)(d) of Regulation (EU) 2024/1689

already requires providers of GPAI models to make publicly available a sufficiently detailed summary about the content used for training, according to a template provided by the AI Office; an explicit link between that summary and the information under Article 14 GDPR would avoid duplication and allow a single document serving both regimes.

7. Data minimisation, synthetic data, anonymisation and pseudonymisation

Synthetic data. The synthetic nature of a dataset does not *ipso facto* make it anonymous. Data generated from real data may retain statistical relationships, regularities and sometimes individual values to an extent that allows re-identification in certain contexts; the qualification depends on the actual properties of the dataset and on the concrete risk of re-identification, not on the technique used to produce it. The Guidelines (Par. 37, 38 and 52) should say so.

Anonymisation and pseudonymisation. Par. 38 and 66(h) present anonymisation and pseudonymisation as different measures, but Par. 52 places them side by side in a way that suggests they are interchangeable. Pseudonymisation reduces risk without taking the data outside the scope of the GDPR; only effective anonymisation, assessed against rigorous criteria, does so. Since Guidelines 02/2026 on anonymisation were adopted at the same plenary session of 7 July 2026 and are also under consultation, the final versions of the two documents should be coordinated, in particular on synthetic data and on the anonymity of trained models. Footnote 18 refers to Guidelines 01/2025 on pseudonymisation in the version for public consultation adopted on 16 January 2025; the reference should be checked and, where appropriate, updated.

Syntax-based filters. The regular expressions mentioned in Par. 38 catch identifiers with a recognisable format, such as telephone or social security numbers, but miss proper names, indirect references and personal data embedded in narrative text, which make up by far the greater part of the personal data in a text corpus. The Guidelines should state this limitation and invite controllers to combine syntax-based filters with named entity recognition and, above all, to measure and document their effectiveness: a filter whose false negative rate is unknown offers only apparent protection.

Lifecycle. Data minimisation should apply to the entire data lifecycle, distinguishing between the raw dataset, the cleaned dataset, the training, evaluation and fine-tuning datasets, intermediate checkpoints, backups and copies transferred downstream. The deletion or anonymisation of unnecessary data under Par. 66(g) is not effective if the data survives, possibly in a different form, in any of these copies.

8. Accuracy, reliability of sources and necessity

Types of inaccuracy. Par. 40 extends the principle of accuracy to the output of the model where it constitutes personal data. This extension calls for a distinction between phenomena with different causes and remedies: inaccuracy already present in the source; data correct at the time of collection and later outdated; errors introduced during cleaning or transformation; incorrect probabilistic inference; hallucination proper, that is, the generation of statements with no basis in the training data. Each involves a different stage of the pipeline and a different corrective measure, and distinguishing them would make more concrete the right to rectification under Article 16, currently the most difficult to implement with respect to generative outputs.

Reliable sources. Par. 42 recommends sources that are official, trustworthy and maintained and the avoidance of secondary or outdated aggregators. The reliability of a source, however, can only be assessed against the purpose: an institutional source may be unsuitable for certain uses, while an informal source may be the very object of the analysis, as for models intended to understand colloquial language. It is suggested that the reference to official sources be replaced by criteria of provenance, verifiability, timeliness and fitness for purpose.

Necessity. Par. 51-52 refer to the existence of an equally effective and less intrusive way of pursuing the purpose, citing synthetic and pseudonymised data. Consistently with paragraph 29 of Guidelines 1/2024, to which footnote 28 refers, it should be specified that a less intrusive alternative is relevant only if it is functionally capable of pursuing the interest declared: the theoretical existence of synthetic data techniques does not in itself make the processing of real data unnecessary, and,

conversely, their possible inadequacy does not relieve the controller of the burden of demonstrating that it has assessed them.

9. Legitimate interest, general-purpose models and structure of the balancing test

Determinacy of the interest. Par. 50 accepts that, for a GPAI model, the precise use may not yet have been decided, and recommends referring to the objective pursued (commercial, public or scientific research; internal or external). The Guidelines should clarify what degree of specificity is sufficient for the interest to be, as required by Par. 49, clearly and precisely articulated and real and present, not speculative; otherwise the open-ended nature of the model and the requirement of a determinate interest remain in tension with no criterion to reconcile them.

Causal proximity. The risks listed in Par. 54 (chilling effect on freedom of expression, identification of data subjects who intended to remain anonymous, profiling, deep-fakes) arise at very different points of the causal chain: some from the collection, some from the training, others from the capabilities of the model or from the use a deployer makes of it. In addition to the likelihood of consequences materialising (Par. 61), ANGIF proposes that the balancing test take into account the causal proximity between the processing under assessment and the feared detriment, and the controller's actual ability to influence it, a criterion the Guidelines themselves adopt in Par. 74 through the formula of "responsibilities, powers and capabilities".

Mandatory measures and mitigating measures. Par. 65 rightly states that mitigating measures should not be confused with measures the controller is legally required to adopt anyway. The list in Par. 66 does not fully respect this distinction: point (b) describes the exclusion of certain data categories and sources as important first of all to comply with data minimisation, and point (g) concerns the deletion of unnecessary data, which follows directly from Article 5(1)(c) and (e). Mandatory measures cannot also count as improving the balance. It is suggested that Par. 66 separate what is required from what is additional, so that minimum compliance cannot be presented as mitigation.

Example 7. In Example 7 the balancing test "may generally be considered as met" by an organisation which, among other measures, excludes content protected by copyright or uses only content for which rights holders have not objected to text and data mining under Directive 2019/790. That condition belongs to a different body of law and says nothing about the risk to data subjects (see Section 12). Moreover, the breadth of the concluding wording lends itself to being read as a safe harbour; the example should end by recalling that the balancing test must in any event be carried out on the specific facts of each case.

10. Data subject rights after training

Anonymity of the model. Par. 60 acknowledges that, given the current technical state of the art, personal data cannot be easily deleted from a trained model. The effectiveness of the rights to rectification, erasure, restriction and objection depends on how this is handled when the data has been removed from the dataset but the model can still reproduce or reconstruct it. The preliminary question is whether the model is anonymous. Opinion 28/2024 has set out the criteria for that assessment; Guidelines 03/2026 do not refer to them on this point. ANGIF requests that the Guidelines state expressly that, where the model is not anonymous, data subject rights are exercisable in respect of the model and not only of the dataset.

Types of memorisation. It would be useful to distinguish between verbatim memorisation, memorisation extractable through attacks, probabilistic reconstruction of information, inference of data never present as such in the corpus and mere coincidence of the output with real data. These are technically different phenomena, and attaching identical legal consequences to them would be both unfair to the controller and ineffective for the data subject.

Machine unlearning. Par. 74(iv) mentions machine unlearning as a possible future alternative to output filters and retraining. Its suitability should be subject to verifiable criteria, translated into reproducible and documented technical tests that an independent expert can repeat with comparable results, capable of distinguishing between a mere reduction in the probability of generating a piece

of information, the suppression of the output through filters and the actual neutralisation of the information in the model's parameters. Only the latter, operating *ex nunc* on the model rather than on its surface, comes close to erasure in the proper sense.

11. Data protection impact assessment, FRIA, data protection by design and security

DPIA. The data protection impact assessment appears only in Par. 34, as an optional measure when relying on the exception to individual information. Scraping for the training of generative models as a rule meets several of the criteria in WP248 rev.01, to which the Guidelines refer in footnotes 1 and 17: large-scale processing, innovative use of technology, matching or combining datasets, possible presence of data concerning vulnerable data subjects and invisibility of the processing. At national level, the list of the kinds of processing operations subject to a data protection impact assessment adopted by the Italian Data Protection Authority (Garante per la protezione dei dati personali) by decision No. 467 of 11 October 2018 leads to the same conclusion. ANGIF requests that the Guidelines state that, in these cases, a data protection impact assessment under Article 35 GDPR is as a rule mandatory, and indicate its minimum content.

FRIA. The fundamental rights impact assessment under Article 27 of Regulation (EU) 2024/1689 does not make the DPIA superfluous. Article 27 applies only to deployers of high-risk AI systems that are bodies governed by public law, private entities providing public services, or deployers of the systems referred to in points 5(b) and (c) of Annex III; it does not cover the development of generative models or the collection of training data. Nor do the assessment and mitigation of systemic risks under Article 55 fill the gap, since they concern risks at a collective scale rather than the position of individual data subjects. At the stage covered by the Guidelines, the DPIA is therefore the only prior assessment of the impact on individuals; and since Article 35(7)(c) GDPR refers to the risks to the rights and freedoms of data subjects, without being confined to data protection, it can and must also address the issues identified in Par. 54, from the chilling effect on freedom of expression to the harm to dignity caused by deep-fakes. The Guidelines should say so, and recall that, where the model is integrated into a high-risk AI system, Article 27(4) of Regulation (EU) 2024/1689 provides that the fundamental rights impact assessment shall complement the DPIA already conducted; the DPIA should therefore be drawn up and documented so that it can be passed on and used downstream.

Data protection by design. Many of the measures in the Guidelines (pre-collection filters, exclusion lists, deduplication, output filters) are typical applications of Article 25 GDPR. Linking them expressly to data protection by design and by default would turn them into architectural requirements of the pipeline rather than compliance tasks documented *a posteriori*.

Security. Security under Article 32 is mentioned only incidentally in Par. 34. The scraping chain is exposed to specific threats: training data poisoning, tampering with datasets, compromise of crawlers, supply chain attacks, injection of content intended to be collected and extraction attacks on the model. These affect not only confidentiality but the integrity and traceability of the dataset, and therefore the very possibility of demonstrating compliance. A dataset whose integrity cannot be proven is a dataset whose lawfulness cannot be proven: the Guidelines should link the security measures under Article 32 expressly to the accountability obligation under Article 5(2).

12. Coordination with the Artificial Intelligence Act and copyright law

Example 7 and the references to robots.txt and ai.txt in Par. 37 and 64 bring copyright and text and data mining into the Guidelines. ANGIF submits that a dedicated section on the coordination between the regimes involved is therefore needed. The lawfulness of processing under the GDPR, the permissibility of text and data mining under Articles 3 and 4 of Directive (EU) 2019/790 and compliance with Regulation (EU) 2024/1689 are autonomous levels: the absence of a reservation of rights does not make lawful the processing of the personal data contained in a work; a legal basis under Article 6 GDPR confers no title as regards intellectual property; and compliance with the Artificial Intelligence Act is not a safe harbour under the GDPR, as confirmed by Article 2(7) of Regulation (EU) 2024/1689, which provides that it shall not affect Union law on the protection of personal data.

The three levels interact, however, through common instruments. Article 53(1)(c) of Regulation (EU) 2024/1689 requires providers of GPAI models to put in place a policy to comply with Union copyright law, in particular to identify and comply with reservations of rights expressed pursuant to Article 4(3) of Directive 2019/790, and point (d) requires the public summary of training content referred to in Section 6. Stating expressly how these provisions relate to the GDPR would prevent technical signals created for one purpose from being given meanings they do not have, and would allow controllers to keep a single, consistent and verifiable body of documentation instead of parallel compliance exercises that end up contradicting one another.

13. Special categories of personal data and the GC and Others judgment

Relevant similarities. Par. 67-75 apply to the incidental and residual collection of special categories of personal data the Judgment of the Court of Justice (Grand Chamber) of 24 September 2019, GC and Others, C-136/17, ECLI:EU:C:2019:773, and its formula "within the framework of his responsibilities, powers and capabilities". Par. 72 rightly warns that the judgment is not a general exemption from Articles 9 and 10. For the same reason, Par. 73(a) should specify which relevant similarities must exist for scraping and training to be regarded as analogous to the activity examined by the Court. Indexing and training are different operations: the former makes already published information retrievable, leaving it where it is; the latter incorporates it into a statistical learning process designed to produce new capabilities. As Par. 71 recalls, in GC and Others the operator's obligations apply downstream, by reason of referencing and following a request by the data subject, whereas the developer of a model acts on the data beforehand and independently of any request.

Article 9(2)(e) GDPR. The Guidelines do not mention Article 9(2)(e) GDPR, the derogation most frequently relied on for data collected online. They should recall that the data must have been manifestly made public by the data subject themselves, and that neither mere online presence nor publication by third parties meets that requirement, with reference to the Judgment of the Court of Justice (Grand Chamber) of 4 July 2023, Meta Platforms and Others v Bundeskartellamt, C-252/21, ECLI:EU:C:2023:537.

Article 10 GDPR. Par. 72 extends the reasoning of GC and Others to Article 10, but the conditions in Par. 73-74 are framed only around the categories in Article 9. Personal data relating to criminal convictions and offences require specific treatment, all the more so since their presence in journalistic and judicial archives accessible online is anything but residual.

Chain of remedies. Deletion from the dataset under Par. 74(ii), if it takes place after training, does not remove the data from the model, and point (iii) then leaves protection essentially to output filters. It is suggested that the Guidelines describe a complete chain of remedies covering the raw dataset, the training dataset, checkpoints, the model, outputs and copies transferred downstream. Par. 75 correctly places the *onus probandi* on the controller; the Guidelines should also indicate what evidence is capable of discharging it.

14. Output monitoring, state of the art and documentation of the supply chain

Constant monitoring. Par. 74(iv) requires the controller using the model in an AI system to constantly monitor the output. In services generating millions of responses a day, this may mean very different things (sampling, automated classification, systematic logging, periodic adversarial testing), with different costs, effectiveness and implications. Output monitoring is also itself a processing operation, involving the data of the system's users and the content of their prompts. The Guidelines should indicate which forms of monitoring are considered adequate and with what safeguards for users.

State of the art. Par. 74(iii), 76 and 77 refer to the state of the art for resistance to privacy attacks and for the regular reassessment of measures. For accountability purposes, the Guidelines should indicate, at least by category, the technical tests suitable for uniform use: membership inference tests, extraction tests, memorisation measurements and verification by means of canary data inserted into the corpus. A reference to European or international technical standards would be useful, provided it is made clear that the presumption of conformity attached to harmonised standards by Article 40 of

Regulation (EU) 2024/1689 concerns the obligations under that Regulation and does not extend to the GDPR, where adherence to a standard remains an element of evidence to be assessed in the specific case.

Dataset provenance record. Par. 30-32 recommend disclosing sources, domain names, URLs and periods of collection and, for datasets obtained from third parties, the controller from whom they were obtained. ANGIF proposes that these indications be developed into a standardised dataset provenance record, documenting for each component its origin, date, the crawler used, the selection criteria, the filters applied, the objections complied with, subsequent transformations and transfers. Such a record would provide the controller with evidence of compliance, data subjects with a basis for targeted requests and downstream parties with the basis for the due diligence referred to in Section 2.

Verifiability of the record. The experience of ANGIF's members as technical experts in judicial proceedings shows that such a record has value only if it is itself verifiable. This requires that it be created at the same time as the operations it describes and not reconstructed *a posteriori*; that its entries be protected against alteration by integrity techniques such as cryptographic hashes of datasets and their versions, time-stamping and retention of crawler logs; and that the chain of custody of the data be traceable without gaps from the source to the model. These are the same safeguards that digital forensics practice requires for electronic evidence, in line with ISO/IEC 27037:2012 on guidelines for the identification, collection, acquisition and preservation of digital evidence, and which the Italian procedural system has recognised since Law No. 48 of 18 March 2008 (legge 18 marzo 2008, n. 48), ratifying and implementing the Council of Europe Convention on Cybercrime. They would enable the supervisory authority and, where appropriate, the judicial authority to verify *ex post* what the controller declares *ex ante*.

Repetition of the assessment. Par. 10 notes that websites change their structure frequently and that scrapers need regular updates, and Par. 76 requires regular verification of the effectiveness of measures. It follows, in ANGIF's view, that the compliance assessment cannot be carried out *una tantum*, but must be repeated whenever there is a substantial change in sources, collection criteria, crawler behaviour, categories of data, objection mechanisms or purposes. For activities of particularly significant size, the Guidelines could encourage periodic technical audits, including by independent parties, of the effectiveness of filters, exclusions, anti-memorisation measures and rights management mechanisms.

15. Terminology, protection of minors and editorial remarks

"Sensitive data". Footnote 17 specifies that "sensitive data" is distinct from special categories of data within the meaning of Article 9, but the term recurs (Par. 14, 37, 74(i)) without it always being clear in which sense. Since the English version is the basis for the translations into the other official languages, it is suggested that a consistent distinction be adopted between special categories of personal data under Article 9, personal data relating to criminal convictions and offences under Article 10 and other high-risk data, such as the financial and location data referred to in Par. 37 and 55.

Minors. Minors are mentioned in Par. 46 (reproducing Article 6(1)(f)), in Par. 37 and 74(i) (exclusion of sites or social networks used mainly by minors) and in Par. 55 and 64 (balancing test and reasonable expectations). Given the practicalasy difficulty of knowing data subjects' age in mass collection, the matter deserves unified treatment, indicating criteria for identifying sources with a high presence of minors, preventive measures, the weight of minority in the balancing test and exclusion mechanisms by default.

Roles. A representation relating the technical function performed, the qualification under Regulation (EU) 2024/1689 and the qualification under the GDPR would be useful, since the final sections move without transition from the controller carrying out the scraping to the controller using the model in an AI system (Par. 74(iii) and (iv)), who in practice often belong to different organisations.

Editorial remarks. The executive summary reads “situations were” instead of “situations where”; in Example 1, “to scrape data from the web to and create data sets” contains a superfluous preposition; in Par. 73(c), “prevent the collection if such data” should read “of such data”; in Par. 74(i), “data concerning vulnerable such as minors” lacks a noun; in Par. 74(iv), “should also be considered is necessary” should read “if necessary”; in Par. 75, the quotation “within the framework of his responsibilities, powers and capabilities” closes without an opening quotation mark. Footnote 22 contains incomplete case-law references: the name of the case Rīgas satiksme and part of the name of case C-621/22 are missing, and the SNCF Connect judgment, C-394/23, lacks the European Case Law Identifier. Footnote 5 refers to “p. 64” of Opinion 28/2024, whereas all other footnotes refer to paragraphs.

Genoa (IT), October 2026

ANGIF
Associazione Nazionale Giuristi Informatici e Forensi
The National President
Avv. Elena Bassoli