CONTRIBUTION TO CONSULTATION
EDPB GUIDELINES 03/2026 ON WEB SCRAPING IN THE CONTEXT OF GENERATIVE AI

MFE-MEDIAFOREUROPE brings together leading media brands across broadcasting, VOD and audiovisual production, with operations in Italy, Spain, Germany, Austria, Portugal and Switzerland. Reaching 220 million viewers and employing approximately 13,000 people, the Group is committed to cultural diversity, media pluralism, accountability and consumer protection. MFE reinvests most of its advertising revenues in European content, including films, series, entertainment and professional news.

A) Review of the relevant data protection issues and proposed amendments

As a preliminary consideration, it is important to highlight a fundamental issue: the consequences of scraping are, in practice, irreversible. Once the model has been trained, personal data cannot easily be erased (as acknowledged by the EDPB itself in its document), and the sheer scale of the phenomenon renders ex post safeguards — information provided and rights exercised after the processing has taken place — essentially illusory, all the more so since the data subject is usually unaware of having been subject to scraping. It follows that compliance with the principles must be capable of being verified ex ante, before collection, and that the only effective safeguards may, upstream, take the form of timely information and a prior right to object. The issues set out below should be read in this light.

1. The exemption from the information obligation on grounds of disproportionate effort

This issue arises in paras. 25, 27-28 and 30-31 and in Example 2: the controller may refrain from providing individual information where this proves impossible or would involve a disproportionate effort (Article 14(5)(b) GDPR), with the balancing exercise being carried out, *inter alia*, on the dataset as a whole (para. 28), whereas the publication of a granular list of domains and URLs is framed as a mere good practice, and only where feasible (para. 31). The risk for MFE is therefore twofold and concrete: its editorial content and the data of its users (e.g. artists under contract) could potentially be “harvested” without any information being provided and, since the list of sources would remain optional, neither the publisher nor the data subjects would be in a position to know whether, and to what extent, the domains have been subject to scraping. Techniques for preventing and blocking scrapers, on the one hand, may not be readily feasible for publishers and, on the other hand, would shift onto the latter a sort of “burden of protection” that is disproportionate compared with the advantage enjoyed by scrapers, who would be able to avoid providing information pursuant to the GDPR. Furthermore, even where the publisher were

able, ex post, to detect scraping activities on its digital assets by means of appropriate monitoring tools, such detection would not extend to the data subjects and would in any event not be decisive, since by that time the data would already have been acquired and could not easily be erased. In addition, the mere publication of a privacy notice on the scraper's website (paras. 30 and 34) would be an inadequate remedy, since the data subject does not know that he or she should look for such information.

2. The central role of legitimate interest and reasonable expectations

The central role played by this legal basis in the processing at issue emerges at several points in the Guidelines under consideration (by way of non-exhaustive example only: paras. 43 and 46-50; paras. 62-64 and Example 4).

On a preliminary analysis, the EDPB appears to afford "privileged treatment" to this legal basis which, as is well known, when applied to the case at hand, comes up against established requirements of transparency and awareness vis-à-vis data subjects. Looking at para. 64 of the document, the Board essentially views favourably the possibility that a user may have a "reasonable expectation" that his or her personal data available online could be re-used by third parties. Such an interpretation of the concept of "reasonable expectation" — the assessment of which is required in any *legitimate interest assessment* (so-called "LIA") — could lead to an abuse of the legal basis in question, with consequences to be assessed in other processing contexts as well. It should also be noted that the average user is hardly aware of the very phenomenon of web scraping; where such awareness is lacking, the factual premise of the "reasonable expectation" — which is itself an evident corollary of that awareness — falls away.

In addition, for any publisher, the risk is that its freely accessible editorial content may, under a permissive reading, become subject to scraping almost by default on the basis of legitimate interest alone, notwithstanding that such content was published for a distinct informational purpose, on the basis of a different legal basis and following far more rigorous assessments of any possible prejudice to the rights and freedoms of the data subjects concerned.

3. The "incidental and residual" collection of special categories of data

This issue arises in paras. 67-77: the EDPB extends the reasoning of the Court of Justice (C-136/17), developed in relation to search engines, so far as to allow residual processing of data under Article 9 on the basis of the test founded on responsibilities, powers and capabilities (paras. 72-74). Journalistic and news content is structurally rich in special categories of personal data under Article 9 GDPR (data concerning health, data concerning sexual orientation) and in personal data relating to criminal convictions and offences under Article 10 GDPR processed for journalistic purposes, so that the mass scraping of such archives would systematically ingest such data; since the absence of special categories of data cannot be guaranteed in large datasets, a broad reading of the notion of "incidental" collection

would end up shielding *scrapers* from the prohibition laid down in Article 9, de facto legitimising the large-scale capture of special categories of data drawn from editorial output, without any control on the part of either the publisher or the data subjects.

4. Proposed amendments

While leaving it to the EDPB to carry out any broader assessment of all the issues raised above, a number of targeted proposals for amendments to the text of the Guidelines under consideration are set out below.

- Given that data cannot easily be removed once a model has been trained, compliance with the principles of the GDPR must be ensured upstream, by means of timely information and a prior right to object, rather than through ex post remedies. This is an aspect that the EDPB should develop further in the Guidelines under consideration.
- It is suggested that consideration be given to extending the personal and material scope of the document, covering, on the one hand, cases other than generative AI alone and, on the other hand, providing for protection that is not limited to initiatives by private entities, thus including scraping attributable to public bodies and clarifying the position of data brokers.
- It is suggested that the publication of an up-to-date, granular and searchable list of sources (domains and URLs) and of the collection dates — which the Guidelines currently describe as a mere good practice, and only where feasible (para. 31) — be made mandatory for the controller (developer/scraper) and, in the case of re-used datasets, for the entity re-using them. Similarly, it would be preferable to confine the exemption under Article 14(5)(b) GDPR solely to certain of the purposes referred to in para. 27 (specifically, "…for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes").
- It is suggested that a rebuttable presumption be introduced whereby, where the source expresses its objection by any effective means (robots.txt, ai.txt, CAPTCHA, terms of use, TDM reservation), data subjects cannot reasonably expect their data to be re-used for training purposes.
- Special categories of data: it would be advisable to interpret the requirement of incidental collection restrictively, given that, in the case of very large datasets, the absence of data falling under Articles 9-10 GDPR could not be guaranteed with adequate certainty.
- It would likewise be advisable to make mandatory (rather than merely optional) a prior objection mechanism at domain or publisher level, technically implementable by means of machine-readable protocols; as regards the technical arrangements for exercising such prior objection, the EDPB itself could identify/propose concrete solutions that are understandable and readily implementable by operators at reasonable cost and within reasonable timeframes, so as not to create disparities between

entities with differing means and technologies and to ensure that such solutions can actually be accepted by the market.

B) Review of the relevant issues from a copyright protection perspective and proposed amendments

1. Introduction

The Guidelines under consideration are expressly situated within the framework of the General Data Protection Regulation and address web scraping exclusively as a personal data processing activity aimed at training generative artificial intelligence models. The perspective chosen by the Board is, by definition, partial, since the mass collection of online content for the training of generative systems engages two autonomous and cumulative legal regimes: that of personal data protection and that of copyright and related rights. The latter regime is neither an alternative to the former nor absorbed by it, since the lawfulness of processing under Article 6 GDPR does not exhaust the question of the lawfulness of the reproduction of works entailed by that processing, which remains subject to separate scrutiny under the text and data mining exception.

Throughout the entire document, copyright is expressly mentioned only once, in Example 7, which concludes the section on the balancing test, where reference is made to the exclusion of protected content and to Directive (EU) 2019/790[1]. Every other part of the text, from the qualification of the parties involved to the processing principles, and from the legal basis of legitimate interest to the rules on special categories of data, is framed within the boundaries of the GDPR. In particular, the Guidelines lack the basic statement that the free accessibility of online content does not equate to its free usability, since copyright protection subsists irrespective of whether the data have been made public by the data subject and irrespective of the existence of a valid legal basis for the processing of any personal data embedded in the work.

This approach gives rise to a specific interpretative risk. The reader might infer that the existence of a legitimate interest under Article 6(1)(f) GDPR, accompanied by the mitigating measures described in paras. 65 and 66, is a sufficient condition for the lawfulness of the scraping operation. That approach is incomplete, since the reproduction and extraction inherent in collection and training affect the exclusive right of reproduction and must find separate justification under the text and data mining regime, namely in Articles 3 and 4 of the 2019 Directive and, in the Italian legal system, in Articles 70-ter and 70-quater of the

[1]Directive (EU) 2019/790, Articles 3 and 4; in the Italian legal system, Articles 70-ter and 70-quater of Law No. 633 of 22 April 1941, introduced by Article 1 of Legislative Decree No. 177 of 8 November 2021, in force since 12 December 2021 (Official Gazette No. 283 of 27 November 2021).

Italian Copyright Law[2]. Since the Guidelines are expressly addressed to private entities, the relevant regime is the general one under Article 4, which is subject to the absence of a reservation of rights, and not the regime under Article 3, which is reserved for scientific research and is not affected by opt-outs.

2. The reference to the TDM exception in Example 7 and its technical imprecision

The only express reference, contained in Example 7, is worded in a technically imprecise manner that weakens its scope. The text classifies as compliant the use of public domain content or of content in respect of which rightholders have not objected to text and data mining as permitted by the 2019 Directive. Three aspects appear problematic. First, the wording juxtaposes two heterogeneous situations — that of the public domain, where no exclusive right exists at all, and that of a protected work which is lawfully mined in the absence of a reservation — thereby creating confusion between the absence of protection and the continued existence of protection subject to an applicable exception.

Secondly, the statement that it is sufficient for rightholders not to have objected underestimates the technical requirement laid down in Article 4(3). In the case of content made publicly available online, the reservation of rights must be expressed in an appropriate manner. The wording of the Guidelines, on the one hand, fails to reflect that qualified standard and, on the other hand, ambiguously suggests a case-by-case burden of objection on the rightholder, whereas the rationale of the provision is that of a general and prior reservation.

Thirdly, the reference omits the requirement of lawful access to the source, which is common to both Article 3 and Article 4 and is transposed into Article 70-quater of the Italian Copyright Law, which permits extraction only **from works or other subject matter to which there is lawful access**[3]. The omission is not neutral, since access to protected content behind authentication or through the circumvention of technological protection measures remains unlawful regardless of whether or not a reservation exists, and can therefore under no circumstances fall within the scope of the exception.

3. Technical measures to object to scraping and reservation of rights under Article 4(3)

The point of greatest convergence between the two regimes lies where the Guidelines address the technical measures by which websites object to collection. In the section on data minimisation, para. 37 recommends that the controller exclude from collection websites that clearly object to scraping by means of robot authentication, robots.txt or ai.txt files and CAPTCHA mechanisms; the same tools re-emerge, from the perspective of reasonable expectations, in para. 64 and in Examples 4 and 5. It should

[2]Article 70-ter of Law No. 633/1941, which transposes Article 3 of Directive (EU) 2019/790 and reserves the exception, which is not subject to any reservation of rights, to research organisations and cultural heritage institutions for the purposes of scientific research.

[3]Article 70-quater of Law No. 633/1941: reproductions and extractions are permitted from works or other subject matter to which there is lawful access, provided that such use has not been expressly reserved by the holders of copyright and related rights or by the holders of databases.

nonetheless be specified that the presence of such measures, where circumvented or overcome, does not in itself render the collection lawful, just as, conversely, their absence does not authorise it. These are precisely the same technical signals which, on the copyright side, constitute the machine-readable reservation capable of disapplying the TDM exception under Article 4(3).

The dual significance of these signals is not, however, made explicit. The Guidelines regard them as relevant solely for the purposes of data minimisation and the data subject's expectations, without ever stating that failure to comply with them constitutes, in and of itself, an infringement of copyright. Yet the Code of Practice for general-purpose AI models expressly commits providers to employ crawlers that read and follow the instructions of the robots.txt protocol[4], and the technical landscape has meanwhile been enriched by further standards, from the X-Robots-Tag header to the TDM Reservation Protocol developed within the W3C.[5] Strengthening protection therefore requires explicit recognition that compliance with such signals simultaneously serves a GDPR compliance function and a copyright compliance function, with the consequence that circumventing them undermines both.

In order to introduce copyright into the framework of the Guidelines, it is necessary to rely on the first condition of the legitimate interest test. Para. 49 recalls, drawing on the Board's guidance, that an interest can be regarded as legitimate only if it is lawful, clearly and precisely articulated, real and present[6]. The lawfulness of the interest is therefore a constituent element, and not a merely ancillary one, of the legal basis. It follows that an interest pursued by means of a scraping activity that disregards a valid reservation of rights, or that otherwise infringes copyright rules, cannot be qualified as lawful, with the consequence that the first condition is not met and the legal basis collapses at the outset, even before, and irrespective of, the balancing test.

4. Reasonable expectations and the overlap with the opt-out rationale

The section on reasonable expectations, in paras. 62 to 64 and in Examples 4 and 5, lends itself to a beneficial alignment with the rationale of the reservation of rights. Example 5 already moves in this direction, since it denies a reasonable expectation of scraping where the platform prohibits collection by means of robots.txt and CAPTCHA and expressly states that it does not allow the use of data for the development of artificial intelligence models. The same factual situation that excludes the expectation under the GDPR constitutes, under copyright law, a reservation capable of disapplying the exception, so that the two regimes reinforce each other.

---

[4]Code of Practice for General-Purpose AI Models (GPAI Code of Practice), published by the European Commission on 10 July 2025, Copyright chapter, which commits signatories to employ crawlers capable of reading and following the instructions expressed by means of the Robot Exclusion Protocol (robots.txt).

[5]On the technical standards for expressing the reservation, see, in addition to robots.txt, the HTTP X-Robots-Tag, the TDM Reservation Protocol developed within the W3C and ai.txt files; the European Commission has initiated a reflection on the shift from the notion of machine readability to that of a technically actionable signal.

[6]Regulation (EU) 2016/679, Article 6(1)(f); see also EDPB, Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, adopted on 8 October 2024.

Example 4 is more delicate, as it recognises a reasonable expectation of scraping where the content is freely accessible, not subject to any prohibition and accompanied by a notice from the platform regarding the possibility of collection. From a copyright perspective, the absence of a reservation does not equate to an authorisation to use the content, but merely to the exception not being disapplied, and this only on condition of lawful access and without prejudice to any applicable licensing arrangements. Moreover, the Guidelines themselves state in para. 45 that the absence or inapplicability of a robots.txt file does not amount to consent under the GDPR. The same caution should be extended to copyright, by clarifying that the absence of a reservation does not constitute an authorisation to exploit the work, in order to prevent Example 4 from being read as implicitly equating free accessibility with free usability.

5. Deepfakes, voice and image: the intersection with related rights

Certain passages of the Guidelines refer to content the collection of which also engages related rights. Para. 54 refers to the generation of deepfakes, Example 6 describes the collection of voice recordings for the development of a voice generation tool, and footnote 4 refers to the joint statement of 23 February 2026 on AI-generated imagery. These cases are not limited to data protection and the protection of dignity, since they involve the rights of performers and, where the recordings constitute phonograms, the rights of the producer[7]. Example 6, in particular, is analysed solely from the perspective of legitimate interest, whereas the collection of recorded voices affects the related rights of performers and, often, phonogram producers' rights.

Since Article 70-quater expressly extends the right to reserve to holders of related rights, the opt-out rules also operate in their favour. Strengthening protection requires the Guidelines to acknowledge, starting from the definitions section, that the scraping of audio and audiovisual content involves related rights and that the TDM reservation may also be invoked by the relevant rightholders, so as not to leave this category of rights unaddressed.

6. The silence on Article 53 of the Artificial Intelligence Act as a systemic gap

The most significant gap in the text is its complete silence on Regulation (EU) 2024/1689. Article 53(1)(c) requires providers of general-purpose AI models to put in place a policy to comply with Union law on copyright and, in particular, to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of the 2019 Directive; point (d) requires the publication of a sufficiently detailed summary of the content used for training, according to the template

[7]Articles 80 et seq. of Law No. 633/1941, on the related rights of performers, in addition to the rights of phonogram producers (Articles 72 et seq.); Article 70-quater expressly extends the right to reserve to holders of related rights.

provided by the AI Office[8]. These obligations are complemented by the Code of Practice of 10 July 2025 (see footnote 4) and by the template for the summary of training content of 24 July 2025[9].

The failure to refer to this body of rules deprives the document of an essential anchor and fosters a fragmented compliance landscape, in which the same conduct is assessed in isolation under the two regimes. The proposal for an express reference to Article 53 therefore constitutes the amendment of greatest systemic significance.

MFE – MEDIAFOREUROPE N.V., 29 October 2026

Transparency Register's ID: 91471238809-21

[8]Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 53(1)(c), which requires providers of general-purpose AI models to put in place a policy to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790, and point (d), on the sufficiently detailed summary of the content used for training.

[9]European Commission, Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models, 24 July 2025, adopted for the purposes of Article 53(1)(d) of Regulation (EU) 2024/1689.