Fedoweb welcomes the opportunity to contribute to the public consultation ("**Public Consultation**") promoted by the European Data Protection Board ("**EDPB**") to collect stakeholders' comments on the Guidelines 03/2026 on web scraping in the context of generative AI on 7 July 2026 ("**Guidelines**").

Fedoweb - the Federation of web operators (*Federazione Operatori Web*) - was created in 2000 to meet the requirements of the on-line market and promote the internet as a means of advertising and communication. Today, Fedoweb brings together the main publishing groups and television broadcasters and the leading web operators and stands out in the Italian market owing to the different media it represents. The main areas in which Fedoweb works relate to changes to business and regulators which the digital revolution has brought about in advertising, communications and media. Privacy, digital platforms for advertising and content distribution, the audience survey metrics, as well as standard rules for all web operators, are some of the issues that Fedoweb continuously monitors. Among other things, Fedoweb is a quotaholder of Audicom Srl, which is the Joint Industry Committee ("**JIC**") that measures, for Italy, the audience of multimedia, editorial and/or advertising content through both the Internet and paper press.

The Public Consultation is of particular interest to our Federation and its members, as it offers a timely and necessary opportunity to clarify the application of Regulation (EU) 2016/679 ("**GDPR**") to the large-scale, often covert, collection of personal data from the internet for artificial intelligence ("**AI**") training purposes. Together with the European Commission's ongoing efforts to modernize the European legal framework in response to the opportunities and challenges posed by AI, the Guidelines can represent an important step toward ensuring that innovation develops in a manner consistent with fundamental rights, European values, and the integrity of the information ecosystem.

While supporting the objectives pursued by the EDPB with the Guidelines, and by the European Commission with the "Digital Omnibus" initiative, Fedoweb believes that neither initiative adequately takes into account the delicate situation of Fedoweb's members, who operate as professional publishers that produce, verify and distribute editorial content under long-standing journalistic and legal obligations. From this vantage point, Fedoweb and its members are, first and foremost, victims of unlicensed and often covert web scraping carried out by third parties, including AI developers, who extract their editorial content at scale without authorization, attribution or compensation. In this context, the current regulatory landscape appears to create a possible misalignment affecting operators in the media and communications sector: while European copyright law enables publishers to license and monetize the use of editorial content for AI training purposes, the interaction with the GDPR and its provisions may in practice prevent, or severely restrict, such uses while leaving unlicensed, undisclosed scraping of the same content largely unaddressed in practice.

Leaving aside the "Digital Omnibus", which is still under discussion and whose direction and outcome remain uncertain, and focusing instead on the Public Consultation, namely the Guidelines, it is important to note that the Guidelines do not take into account the position of parties, such as Fedoweb's members, who are subject to scraping and are, in practice, unable to limit it in any way except through direct agreements providing for the sharing of their articles with AI system providers. In this context, the Guidelines do not address a broader and increasingly relevant scenario, namely the possibility for AI systems to train on high-quality editorial content that may inherently contain special categories of personal data, and personal data relating to criminal convictions and offences ("**judicial data**"), that cannot be removed or anonymized given the nature of editorial content: anonymization is not only substantially impossible (at most, what can be achieved is pseudonymization) but is also fundamentally incompatible with the informative nature of the editorial content itself. The absence of a clear and workable legal basis for such processing risks preventing AI systems from relying on authoritative journalistic sources and could therefore have significant repercussions not only for economic operators, including the media sector, but also for society at large.

More broadly, this regulatory friction raises policy concerns that extend beyond the media sector alone. It risks undermining the economic sustainability of the European media ecosystem, limiting the ability of AI systems

to rely on trusted and authoritative sources of information, and ultimately affecting the quality and reliability of information circulating within AI-driven environments. These dynamics may in turn have wider implications for media pluralism, democratic resilience, and the integrity of public discourse in the European Union

Against this background, and while the “Digital Omnibus” is still ongoing with no certainty as to its timing or final outcome, the Guidelines represent a timely and concrete opportunity to clarify the interaction between the relevant legal frameworks and to identify solutions capable of safeguarding freedom of information.

Fedoweb wishes to stress, from the outset, that it is not seeking a general derogation from the GDPR. Its concern is narrower and more specific: that data protection rules, as currently applied, risk making a lawful, transparent and auditable editorial content licensing market unworkable, while unauthorized, covert scraping of the same content continues largely unchecked. In this respect, the Guidelines appear, in practice, to implicitly accept generalized scraping, provided that the controller - in this case, the scraper - adopts certain measures without, however, addressing the position of those who hold proprietary rights over the content being scraped. In Fedoweb's view, the most effective way to protect such rightsholders is precisely to enable and facilitate the conclusion of licensing agreements between publishers and AI developers, rather than leaving the legitimacy of large-scale scraping to be assessed solely by reference to the measures adopted by the scraper itself.

The following contribution highlights the key legal and practical challenges arising from the current framework, explains their implications for the sustainability of the media sector and the quality of information available through AI systems, and proposes possible avenues for regulatory clarification that could help address these concerns.

This contribution is made without prejudice to the position and/or contribution that may be taken independently by any Fedoweb’s members, which are therefore free to propose additional and/or different views.

*

1. Copyright law, data protection rules and AI training and development.

The development of AI systems increasingly relies on large volumes of high-quality data. In this context, professionally produced editorial content represents a particularly valuable source of information for AI training, as it combines linguistic quality, editorial oversight, and factual verification.

At the same time, the European copyright framework enables publishers to license and monetize the use of their editorial content for a variety of purposes, without excluding AI training. Such licensing arrangements represent an important opportunity for publishers to generate revenues from the use of their content in the evolving AI ecosystem.

However, editorial content - particularly news reporting - frequently contains personal data relating to identifiable individuals. In many cases, journalistic reporting may also include special categories of personal data, such as political opinions, religious beliefs, health-related information, or references to criminal proceedings.

As a consequence, the use of editorial content for AI training purposes may trigger the application of several provisions of the GDPR that significantly constrain such processing.

In particular:

- consent under Article 6(1)(a) GDPR is not a viable legal basis in this context. Journalistic archives often contain references to thousands or even millions of individuals across decades of reporting, many of whom are no longer identifiable or contactable. This challenge is further compounded when considering real-time sharing of newly published content for AI training purposes;

- the applicability of legitimate interest under Article 6(1)(f) GDPR remains highly uncertain in the context of large-scale AI training, especially given the complexity and scope of journalistic datasets;
- the processing of special categories of personal data under Article 9 GDPR adds and additional layer of complexity. News reporting frequently involves matters of public interest that inherently include such information; however, the derogation under Article 9(2)(g) GDPR does not appear to provide a suitable legal basis for the processing of such data for AI training purposes;
- Article 10 GDPR imposes additional constraints when journalistic content refers to criminal proceedings, further limiting the legal pathways for using such content in AI models. The EDPB Guidelines themselves confirm the scale of this impasse: special category data processing remains prohibited in principle, and the "responsibilities, powers and capabilities" test derived from the Court of Justice's case-law is framed only as a basis for tolerating the *incidental and residual* collection of such data during scraping, where the controller cannot reasonably prevent it. It is not, and does not purport to be, a legal basis for the **intentional and purposeful** use of editorial content that a publisher has deliberately selected, verified, and offered for licensing precisely because it may contain information relevant to matters of public interest, including special category or judicial data. Publishers therefore find themselves in a paradox: an entity that scrapes their content covertly and indiscriminately may, within limits, invoke a residual-collection rationale, while a publisher who seeks to license the same content transparently, under contract and with full knowledge of its contents, has no equivalent pathway available;
- similarly, the conditions identified in the Guidelines by reference to the case law of the Court of Justice of the European Union may not be applicable to the case at hand. In particular, those conditions appear to address situations in which special categories of personal data are only incidentally included in the scraped content. By contrast, news and current affairs articles are likely to contain such data on a regular basis, if not almost invariably - political opinions being an obvious example. This is all the more relevant in light of the Court's subsequent case law, which has extended the concept of special categories of personal data to encompass data inferred from other information. In any event, the case law relied upon in the Guidelines does not address judicial data, which are also routinely included in editorial content;
- the derogations related to journalistic activities may not be sufficient to allow the training of AI systems with journalistic contents, even when the training data consists of lawfully published articles on topics of public interest and publicly accessible by anyone.

In addition to these legal constraints, important technical limitations must also be considered. Technologies such as machine unlearning - which would allow AI systems to selectively remove specific information from trained models - are still at an early stage of development and are not yet widely available. As a result, ensuring full compliance with certain data subject rights, such as the right to erasure or the right to object, may prove extremely difficult once personal data has been incorporated into training datasets.

Taken together, these legal and technical constraints create a situation in which **publishers may be legally entitled under copyright law to license their content for AI training purposes, but be prevented by doing so due to data protection restrictions**.

This structural misalignment risks discouraging lawful licensing arrangements between publishers and AI developers, thereby weakening the economic sustainability of the media sector while simultaneously limiting the availability of trusted editorial sources for the training of AI systems.

2. Authorized licensed access, as opposed to covert mass scraping.

This misalignment is made more acute by the fact that the current framework, in practice, risks treating authorized and transparent licensing less favorably than unauthorized, covert scraping. The Guidelines

themselves describe web scraping as a technique that extracts large amounts of data, often without data subjects being aware of it, and acknowledge that it can be difficult for a scraping controller to know, at the point of collection, what personal data - including special categories of data - its activity has captured.

A publisher that enters into a licensing agreement with an AI developer operates on a fundamentally different basis. The content is identified, its provenance is known, the licensor is an accountable and identifiable legal entity, the terms of use are contractually defined, and the processing can be made subject to audit, logging and purpose limitation from the outset. By contrast, mass or indiscriminate scraping of the same editorial content - frequently carried out in disregard of technical opt-out signals such as robots.txt or ai.txt files, and without any contractual relationship or transparency towards either the publisher or the data subjects concerned - involves none of these safeguards.

Fedoweb considers that any regulatory clarification in this area should expressly distinguish between these two scenarios and should not, even inadvertently, place a greater regulatory burden on disclosed, contractual and auditable licensing arrangements than on undisclosed scraping. A framework that makes licensing legally more burdensome than unlicensed scraping would perversely reward the latter and discourage the very practices - transparency, accountability, traceability - that the GDPR is designed to promote. Controlled and transparent access models, where the publisher authorizes and documents the use made of its content, should be recognized and rewarded as the preferable and lower-risk pathway, including for the purposes of any proportionality or balance assessment under Articles 6, 9 and 10 GDPR.

3. The growing role of AI systems as gateways to information.

The legal uncertainty described above must also be assessed in light of a broader structural transformation in the way citizens access and consume information online.

Generative AI systems are increasingly acting as intermediaries between users and information. Rather than accessing information directly through websites or news outlets, users are progressively turning to AI assistants, search-based generative tools and conversational systems to obtain summaries, explanations and answers – and also journalistic contents.

Recent research highlights the speed at which this transformation is occurring. The use of generative AI tools to access and process news and information is rapidly expanding, particularly among younger audiences. The trend clearly indicates that AI systems are progressively becoming an important gateway through which citizens access information.

This shift is particularly evident among younger generations. Data from **Eurostat** show that the use of generative AI tools is significantly higher among individuals aged 16–24 compared to older age groups[1], suggesting that AI-assisted information retrieval is likely to play an increasingly central role in future information ecosystems.

As AI systems progressively mediate access to information, the **quality and reliability of the sources used to train these systems become critically important**.

Professionally produced editorial content represents one of the most reliable sources of information available in the online environment. Journalistic content is generally produced by trained professionals, subject to editorial review and to professional and ethical standards governing journalistic conduct. While these standards may vary across Member States, they share an emphasis on accountability that distinguishes professional journalism from unverified online sources.

[1] Recent data show that, in 2025, nearly two-thirds (63.8%) of young Europeans (aged 16-24) used generative AI tools. The study is available at https://ec.europa.eu/eurostat/web/products-eurostat-news/w/edn-20260210-1 .

Editorial content therefore combines linguistic quality, factual verification and institutional accountability. These characteristics are relevant not only to the quality and reliability of AI outputs, but also to the assessment of the purposes and safeguards applicable to the processing of any special categories of personal data or data relating to judicial matters contained in such content. In particular, content obtained under a transparent licensing arrangement with an identifiable and accountable publisher, capable of providing information on its provenance, context and the lawfulness of its original publication, presents a materially different risk profile from data collected through indiscriminate scraping of unverified or anonymous sources. This distinction should be taken into account when assessing the necessity and proportionality of the processing and the safeguards appropriate to the specific context.

Ensuring that AI systems can rely on trusted editorial sources during training is therefore of paramount importance: not only to improve the quality and reliability of AI outputs, but also to reduce the risk of misinformation, inaccuracies and misleading content being amplified through AI-generated responses.

Conversely, if regulatory constraints prevent AI developers from lawfully accessing and using authoritative editorial content - e.g. through licensing arrangements with publishers - AI systems may increasingly rely on less reliable online material, including unverified or low-quality sources obtained through indiscriminate web scraping. Such a dynamic could have significant consequences for the quality of information made available to users and, more broadly, for the integrity of the digital information ecosystem.

For these reasons, enabling the responsible and transparent use of high-quality editorial content for AI training purposes should not be viewed solely as an economic matter for the media sector. Rather, it should be recognized as a broader policy priority, closely linked to safeguarding information integrity, media pluralism and democratic resilience within the European Union.

This is ultimately a question of balance between fundamental rights and public interests that are all recognized under EU law: the protection of personal data, on one hand, and press freedom, information pluralism and the sustainability of an independent media sector, on the other. None of these interests should be treated as automatically prevailing over the others.

Fedoweb does not suggest that data protection should yield to commercial or editorial interests; rather, it submits that a workable balance requires a legal pathway through which publishers can lawfully license editorial content - including content incidentally containing special category or judicial data - under proportionate safeguards, rather than a regime that, in practice, forecloses that pathway while doing little to prevent the same content being obtained through unlicensed scraping.

4. Economic implications for the sustainability of the European media ecosystem.

The framework briefly described above has direct economic and societal consequences, particularly for publishers and media operators across the European Union.

As AI systems increasingly act as the primary gateway to information, audiences may consume content directly through AI outputs rather than visiting news websites or subscribing to traditional media outlets. While this shift can offer convenience for users, it reduces traffic and engagement for publishers, which in turn undermines advertising revenues – i.e. the primary source of funding for the great majority of online news services in Europe.

In this context, the ability of media operators to license editorial content for AI training purposes represents a critical pathway to monetize their work. European copyright law allows publishers to grant such licenses, creating a legal and sustainable framework to generate revenue while supporting AI development. However, as already highlighted, GDPR constraints - particularly regarding special categories of personal data and judicial data in journalistic content - may severely restrict such licensing arrangements.

The undesirable consequences can be:

- **unlicensed scraping of content by AI providers**, which risks undermining the economic viability of professional journalism. This practice offers no compensation to the original content creators and bypasses editorial accountability mechanisms; and
- **reduced reliance on high-quality editorial content for AI training**, which may lead AI systems to learn from unverified or lower-quality online sources. This, in turn, affects the accuracy and reliability of AI-generated outputs consumed by the public.

This dual risk poses a long-term threat to the **sustainability of European media outlets**, the **plurality of information sources**, and the **quality of public discourse**. Younger audiences, who are the most frequent users of generative AI tools, are particularly exposed.

To safeguard both the economic sustainability of the media sector and the integrity of AI-driven information ecosystems, **it is essential to enable publishers to license editorial content for AI training under clear, proportionate, and predictable legal conditions**. Such a solution would:

- allow publishers to monetize their content in a manner consistent with copyright law;
- reduce unregulated scraping of content by third-party AI providers;
- ensure AI systems are trained on reliable, authoritative sources, improving output quality;
- support media pluralism and the availability of diverse, high-quality journalism across Europe.

In other words, providing a clear legal basis (or extending any derogations for journalistic activity) for responsible sharing of editorial content would create a **win-win scenario**, simultaneously fostering innovation in AI, protecting fundamental rights and privacy, and preserving the economic and democratic foundations of the European media ecosystem.

5. AI Training, Special Categories of Personal Data and Judicial Data.

A critical challenge for European media operators arises from the use of editorial content to train AI systems where such content contains special categories of personal data within the meaning of Article 9 GDPR, or personal data relating to criminal convictions and offences under Article 10 GDPR. While copyright law may allow publishers to license their content for AI training, such licensing does not in itself provide a lawful basis for processing personal data, which remains subject to the requirements and restrictions of the GDPR.

The Guidelines identify only limited circumstances in which AI providers may process special categories of personal data, namely where such data are processed incidentally and subject to data minimization measures. However, these conditions appear difficult to reconcile with the nature and function of journalistic content. In particular, they may effectively require sensitive information to be removed or rendered anonymous to an extent that is incompatible with the editorial content's informative purpose. News reporting frequently relies on information concerning individuals' political opinions, health, or other sensitive characteristics precisely because such information is relevant to matters of public interest: removing that information may deprive the content of essential context and meaning, thereby undermining its value as a source for training AI systems. Even where anonymisation is technically feasible, it may materially alter the content and impair its ability to fulfil its original informative function.

A further difficulty concerns data relating to criminal convictions and offences under Article 10 GDPR. The pathways identified in the Guidelines for the processing of special categories of personal data do not expressly address such data, leaving unresolved how AI providers may lawfully process judicial information contained in editorial content for training purposes. As a result, publishers may be able to license their content from a copyright perspective while remaining unable to establish whether, and under what conditions, the personal data embedded in that content may lawfully be processed for AI training. This creates significant legal

uncertainty for publishers seeking to make professionally produced editorial content available for the development of AI systems.

As a result, AI systems risk being trained on less authoritative sources, increasing the likelihood of misinformation, fake news and lower-quality outputs that may influence public opinion and undermine trust in both news and AI-generated content. This gap is particularly concerning given the rising reliance on AI for news consumption, noted above.

If AI systems cannot lawfully access trusted editorial content, the economic sustainability of media outlets and the plurality of information are at clear risk. Moreover, continued unlicensed scraping may persist, weakening copyright protections and depriving publishers of the revenue necessary to maintain high-quality journalism.

6. Proposed Pathway.

To address the legal and operational challenges outlined above, the Guidelines should explicitly allow the processing of editorial content containing special categories of personal data, as well as judicial data, for AI training purposes. Such a provision should aim to reconcile multiple objectives: ensuring compliance with data protection rules, respecting copyright, supporting media sustainability, and fostering innovation in AI.

By providing a clear and harmonized pathway, the EU could balance the needs of AI developers with the economic viability of media operators, while safeguarding pluralism and the integrity of information.

In practice, this approach could include several key elements.

First, the scope of the exemption should explicitly cover editorial content created and published by media operators, including articles, reports, and other journalistic material that may contain sensitive or judicial data. Publishers should be authorized both to share such content with AI providers for the purpose of training AI systems and to monetize this sharing, ensuring that the economic value of editorial work is preserved. AI providers, in turn, should be allowed to use this content to enhance language models, verify facts, and improve the quality and reliability of AI-generated outputs.

Second, the framework should support media sustainability and copyright enforcement. By explicitly permitting publishers to license content under commercial agreements, the regulation would help maintain essential revenue streams while providing legal certainty that discourages unlicensed scraping. This, in turn, would facilitate the creation of high-quality AI datasets in a lawful and controlled manner, benefiting both the media sector and AI development and, consequently, the society at large.

Third, strong safeguards for personal data must remain central to any such framework. Fedoweb is not proposing a general derogation from the GDPR, but a set of proportionate technical, organizational and contractual safeguards tailored to licensed editorial content, in place of a blanket prohibition that applies equally to licensed and unlicensed uses alike.

Processing should be strictly limited to AI training purposes and should not be extended to unrelated profiling or commercial exploitation. At the same time, it should be explicitly recognized that measures such as data minimization or pseudonymization should not be mandatory when they would prevent the sharing of editorial content in real time, or undermine the reliability and completeness of editorial content. Imposing such restrictions could compromise both the quality of the content and the ability of AI systems to provide accurate, high-quality information, particularly for news and current affairs, undermining the very purpose of training AI on trusted editorial sources.

The above is proposed as a proportionate alternative to a blanket prohibition on the processing of special category and judicial data contained in licensed editorial content - not as a substitute for compliance with the GDPR, but as the practical expression of that compliance in a licensing context.

Finally, alignment with GDPR principles should be ensured as well. The processing should be explicitly recognized as compatible with legitimate journalistic interests, and consent from individuals mentioned in the content should therefore not be required.

Overall, such a framework would create a balanced and future-proof legal environment, enabling AI systems to benefit from authoritative editorial sources while ensuring the protection of personal data, the sustainability of media operators, and the integrity of public information.

*

To conclude: the Guidelines offer a significant opportunity to resolve a possible misalignment between copyright law, data protection, and AI development.

Without targeted clarification, publishers may not lawfully license editorial content, threatening media sustainability and encouraging AI systems to rely on lower-quality sources. Conversely, by enabling the responsible and safeguarded use of editorial content for AI training, the European Union can simultaneously promote innovation, ensure high-quality information, support media pluralism, and strengthen democratic resilience.

October 30, 2026

Giancarlo Vergori

Fedoweb President

