Observatoire ARGA

INTERNATIONAL ANALYTICAL REPORT

Study #21

Generative Artificial Intelligence and Web Scraping

Legality of data collection, transparency of training datasets, and developer responsibility

Organization: Observatoire ARGA

English version

21 rue de l'Aviation, 64600 Anglet, France

info@argaobservatory.org | www.argaobservatory.org

October 22, 2026

Contents

ANNOTATION.... 4
1. Subject, purpose of the study and the theses being tested.... 4
2. Conceptual boundaries: crawling, scraping, mining, and model training.... 5
3. Data lifecycle map and delineation of operations.... 6
4. Hierarchy of sources of law and the status of EDPB Guidelines 03/2026.... 7
5. When does the GDPR apply to public internet data?.... 8
6. Controller, processor, joint controller and seller of data.... 9
7. Availability of the page and the legality of its further use.... 10
8. Limiting the Goals: A Research Formula Instead of the "AI Training" Slogan.... 10
9. Legitimate interest: the criterion of certainty and legitimacy of the goal.... 11
10. Legitimate interest: the criterion of necessity and alternatives.... 12
11. Legitimate interest: balance of consequences and expectations of the subject.... 13
12. Other grounds for processing, consent and false universal solutions.... 14
13. Special categories of data: Article 9 of the GDPR.... 14
14. Information on convictions and offenses: separate article 10.... 15
15. Data of minors and vulnerable persons.... 16
16. Transparency requirement and Article 14 of the GDPR when data is obtained indirectly.... 17
17. Disproportionate effort criterion: the exception must not swallow up the rule.... 18
18. Minimization: Before, During, and After Extraction.... 18
19. Reliability, time and the problem of outdated information.... 19
20. Data protection impact assessment and risk documentation.... 20
21. Right of access, rectification, erasure and restriction of processing.... 21
22. Objections to processing and taking into account website restrictions.... 22
23. Model weights, memorization and the question of "forgetting".... 22
24. Anonymization, pseudonymization and re-identification.... 23
25. Origin Catalog: Verifiable Evidence for Each Source.... 24
26. International transfers and third country access.... 25
27. Dataset security and contractor responsibilities.... 25
28. Intellectual property: a parallel, not an alternative, legal regime.... 26
29. AI Act: GPAI Documentation, Copyright, and Public Summary.... 26
30. The French approach: CNIL and safeguards for online collection.... 27
31. UK: ICO position and comparative restrictions.... 28
32. Opaque kit suppliers: due diligence before purchasing.... 28
33. Public reporting: detailed enough, but not dangerous.... 29
34. Independent audit: subject of verification and limits of conclusion.... 30
35. Statistical sampling: why a few successful tests do not prove safety.... 30
36. Filter performance metrics and classification error handling.... 31
37. Sanctions compliance: risk of secondary dissemination of false identity.... 32
38. International search, extradition and criminal proceedings data.... 33
39. Suppliers and developers: responsibility for decisions with incomplete information.... 34
40. Proposals of the Observatoire ARGA for EDPB practice and international dialogue.... 34
41. Conclusions, limitations, and future research agenda.... 35
Appendix 1. Terminology map for independent opinion.... 36
Appendix 2. Matrix of legal grounds and additional restrictions.... 36
Appendix 3. Form LIA: Specification of Legitimate Interest.... 37
Appendix 4. LIA Form: Necessity and Less Intrusive Alternatives.... 38
Appendix 5. Balance of Interests: Model of Harm Caused and Prevented.... 38
Appendix 6. Admission and prohibition of Internet sources.... 39
Appendix 7. Source Intake Sheet: Primary Source Card.... 39
Appendix 8. Dataset Manifest: Minimal Structure of Proof of Origin.... 40

Приложение 9. Проверка поставщика уже собранного корпуса....41
Appendix 10. RACI and delineation of powers of participants....42
Appendix 11. Public information notice under Article 14 of the GDPR: draft structure....42
Appendix 12. Justification for Article 14(5)(b): Documented Procedure....43
Приложение 13. DPIA: карта угроз правам и свободам....44
Appendix 14. Stop rule for data under Article 9 of the GDPR....44
Appendix 15. Stop Rule for Information on Criminal Offenses....45
Appendix 16. Child Protection: Domain and Content Exception Rules....45
Appendix 17. Procedure for processing a data subject's request....45
Appendix 18. Map of removal execution at different levels of the model....46
Appendix 19. Audit protocol of actual crawler operation....47
Appendix 20. Data protection model in the contractor chain....48
Appendix 21. Copyright and Text and Data Mining Rights: Separate Check....48
Appendix 22. The AI Act, GDPR, and Disclosure of Training Content....48
Appendix 23. Map of cross-border access to personal data....49
Appendix 24. Plan for statistical verification of source prohibitions....49
Appendix 25. Personal Data Detection Test Matrix....50
Appendix 26. Independent audit program for 20 working procedures....51
Appendix 27. Evidence Chain: example of movement of one document....52
Appendix 28. Case Study: Wrong Sanctions Attribution....52
Appendix 29. Case Study: Outdated Information on Criminal Prosecution....53
Appendix 30. Case Study: Broker Sells an "Anonymous" Set....53
Appendix 31. Public Transparency Report: Minimum Indicators....54
Appendix 32. Draft proposals to EDPB Guidelines 03/2026....54
Appendix 33. 90-day plan for the implementation of organizational methodology....55
Appendix 34. Criteria for acceptance of the report by an independent reviewer....56
Appendix 35. Register of legal and technical issues for subsequent review....56
Appendix 36. Designing a crawler as a legally controlled system....57
Appendix 37. Judicial and evidentiary assessment of the origin of the disputed fragment....58
Appendix 38. Depersonalization, linking, and the risk of reverse attribution....59
Appendix 39. Managing data deletion from backups and training epochs....60
Appendix 40. Procedure for challenging the decision of an independent auditor....61
Appendix 41. Designing a Clear Scraping Notification: A Test with a Real User....62
Appendix 42. Distinction between private scraping and law enforcement access to data....63
Appendix 43. Pre-training, retraining, retrieval and processing of user queries....64
Appendix 44. Method of Responding to Common Unproven Legal Assertions....65
Appendix 45. Case study of negotiations with a broker: a detailed procedure for evidence-based procurement....65
Appendix 46. Justification for disproportionate effort in individual notification....67
Appendix 47. National Legal Restrictions on Public Registers: Conflict Map....68
Appendix 48. Statistical comparison of the quality of two filtration systems....69
Appendix 49. Incident during the opening of the training corpus: a combination of two procedures....69
Appendix 50. Educational file: request for deletion due to name overlap....70
Sources and regulatory and methodological materials....72

ANNOTATION

Study No. 21 examines the legal limitations of automated extraction of publicly available internet data for developing generative models. The analysis distinguishes between page availability, the legal basis for processing personal data, the lawful use of content, and the evidentiary responsibility of developers. It examines the provisions of the GDPR, Article 53 of the EU AI Act, recommendations from the CNIL and ICO, the position of the EDPB, and practical methods of independent verification.

The central source is EDPB Guidelines 03/2026 on web scraping in the context of generative AI. Version 1.0 was accepted for public consultation on 7 July 2026, published on 8 July, and the consultation period continues until 30 October 2026. As of the date of this report, the draft provisions do not constitute the final version of the recommendations following the consultation. Observatoire ARGA's proposals are separate from the mandatory requirements of EU law [2][13].

The practical section includes 50 appendices: forms for legitimate interest analysis, DPIAs, source control, data broker verification, website restrictions and subject rights, provenance manifestos, independent audit procedures, and simulated case studies. The report does not contain audit results that are not available to internal developers and does not attribute violations to any organization without reviewed evidence.

Key findings

Position	Methodological conclusion
Legal prerequisite	The public availability of the material does not replace the grounds of Article 6 of the GDPR, the provisions of Articles 9/10 and the rights of subjects.
Independent verification	Sources, targets, roles, actual logs, and filter effectiveness are assessed, not just the presence of declarations.
Special risk	Training on data on criminal records, health, children, and false positives requires separate constraints.
Remediation	Removing from the set and removing the influence of an already trained model are different technical tasks.
EDPB Status	Guidelines 03/2026 as of the date of the study is a draft for public comment and not a stand-alone new law.

1. Subject, purpose of the study and the theses being tested

This study examines the legal status of automated data collection from open internet sources for training, retraining, and evaluating generative models. The central question is not whether a program can technically load a page, but rather under what conditions an organization has the right to extract, combine, store, and reuse information about specific individuals. The automatic availability of material does not make it a "nobody's" resource, nor does it make it data free from legal restrictions. The study's starting point is to distinguish between technical availability, the legality of access, the legality of personal data processing, and the permissibility of using protected works [1][2][11].

This report was prepared as a normative and methodological study as of October 22, 2026. It makes no claim that the authors accessed internal training datasets of specific companies, conducted

forensic analysis of other companies' models, or established any violation of the law by any named developer. The case studies are explicitly identified as simulated situations; the forms and thresholds provided are proposals of Observatoire ARGA, not official decisions of the supervisory authority.

The research hypothesis being tested is as follows: transparency of the origin of datasets, documentation of the three-part legitimate interest test, and technical evidence of minimization can reduce the risk of infringement, but the existence of a document, robots.txt policy, or domain exclusion alone is not sufficient evidence of the lawfulness of all stages of processing. Compliance must assess the legal basis, the actual procedure, and the actual effect on the data subject as a whole [1][2][3].

For practical analysis, the following chain is used: goal formulation → source selection → automatic extraction acceptability → filtering → inventory → model training → personal data replication verification → request and objection processing → deletion or restriction of further use. At each link, the decision owner, evidentiary artifact, applicable rule, and the possibility of independent verification are recorded. The inability to confirm one key transition cannot be compensated for by a large volume of general declarative documentation.

Research question	The result being tested
Is there a legal basis for the specific processing?	Recording of the purpose, the applicable GDPR article and the materials of its assessment
Can unnecessary data be excluded before collection?	List of rules for selecting domains, prohibitions, and filters
Can the subject exercise rights in practice?	Available request channel and execution log
Is the origin of the data reproducible?	Linked versions of the registry of sources, collections and models
Does an independent expert confirm the result?	Limited in scope and evidence audit report

2. Conceptual boundaries: crawling, scraping, mining, and model training

Web crawling refers to the automated discovery of network resources and the following of the links found. Web scraping is the extraction and storage of content or its structured elements. Targeted sampling based on permitted categories differs significantly in the scale of risk from unrestricted web crawling. In Guidelines 03/2026, the EDPB specifically distinguishes between targeted and untargeted scraping and considers collection from static and dynamic pages [2].

Text and data mining refers to the analytical processing of digital material to identify information, including patterns, regularities, and correlations. It is not synonymous with the legality of collection, and a special copyright exception does not override the requirements for protecting personal data. A training dataset differs from a raw capture: it may contain sanitized documents, fragments, metadata, embeddings, annotations, and derived features. Each operation may raise independent legal issues [1][11].

The terms "public page," "open source," and "access without an account" describe the conditions of acquisition, but in themselves say nothing about permissible subsequent use. Even a searchable publication may contain information about a private individual's address, illness, criminal

prosecution, minor, or private correspondence. Uploading and structuring such information involves the processing of personal data if the individual is identified or identifiable [1].

A distinction must be made between a model, which can be explored through queries, and a training dataset, which is fully accessible. An external memorization test demonstrates a specific behavior pattern under specified conditions and does not establish the composition of all learning sources. Similarly, the absence of a match does not confirm that the disputed document never existed in the training dataset. The methodology should explicitly state the limits of knowledge and not confuse statistical testing with a complete reconstruction of the data.

3. Data lifecycle map and delineation of operations

Legal assessment begins before the crawler is launched. First, the interest and specific goal are defined: for example, multilingual training of a text model for a search interface. Then, the data required to achieve this goal is described, and less intrusive alternatives are compared. Prior to technical collection, records must exist of the selected domains, content categories, legal risks, retention periods, and expected subsequent transformations [1][2].

After collection, data typically undergoes cleansing, de-duplication, language-based segmentation, quality assessment, annotation, and movement between computing service providers. It is a mistake to treat these steps as a single, indifferent technical operation: the typical risk arises when re-merging a dataset originally limited to one purpose with another database that allows for identification of individuals, or when storing unused copies in backup storage. Legal assessments must consider the entire lifecycle [1][2].

Each dataset maintains a versioned manifest with a unique identifier, date, source category, processing reason, responsible department, attributes of possible special categories, legal status, access mode, and links to related model versions. The manifest itself does not prove legality, but it allows for the reconstruction of decision history, identification of gaps, and the fulfillment of deletion requests without blindness.

During the training phase, it's important to distinguish between cleaning up training examples and changing model weights. Removing a row from the catalog does not eliminate its impact on an already trained model. The approach to correction is determined by the degree of risk, technical feasibility, evidence of retention, the availability of subsequent training, and the feasibility of inferring or updating the model. The internal decision is recorded, indicating the remaining risk and the rationale [3][10].

Stage	Risk	Control artifact
Selection of sources	Unlawful collection, other people's prohibitions	Register of categories and admission to collection
Extraction	Accidental theft of private data	Crawler logs, hashes, exceptions
Cleaning	Preserving residual identifiers	Filtering protocol and testing
Unification	Unexpected re-identification	Map of connections between sets and DPIA
Education	Memorisation, information leak	Model card, extraction test
Release	Difficulty in implementing an objection	Complaint channel, correction plan

A practical reconstruction of data movement must begin with the control date. The researcher records which crawler was used, what source rules were set, what code was used for scraping, and how the object ended up in the training set. Without logs, it is impossible to reconstruct the history retroactively based on the current system configuration. Current collection restrictions do not prove that the same policy was in effect when the disputed material was obtained. For the same reason, the audit report records version identifiers as evidence, not as convenient internal labels. A minimum sufficient set includes information on the extraction date, origin, filtering algorithm, transitions between stages, and the point of transfer to the contractor. Operational history is especially important for legal or administrative assessments, when the operator claims to have corrected the data after a complaint, but does not demonstrate whether a model trained on the previous corpus was released.

The issue of temporal inconsistency is considered separately. A record could have been legally published at one point, then corrected or retracted, but still circulate in the archive, search cache, and derived dataset. Validity control cannot be reduced to a one-time "source public" label. For sensitive categories, repeated status checks are useful, especially before each training epoch or before transferring the collection to another operator. The proposed control consists of five questions: does the source currently exist, does it correspond to the current source document, does it belong to an excluded category, are the legal restrictions preserved, and is there a reproducible path to update or delete. This model does not impose an unlimited obligation on the developer to check the entire internet daily, but it does create criteria for proportionate control for known-significant collections.

4. Hierarchy of sources of law and the status of EDPB Guidelines 03/2026

The legal basis in the EU is the GDPR (Regulation 2016/679), which applies directly to the conditions of its territorial and material scope. With regard to the use of generative models, the provisions of the AI Act complement the picture, including the obligations of GPAI model providers regarding documentation, copyright compliance policies, and public descriptions of educational content; these obligations cannot be construed as a blanket permission for any data processing [1][11][12].

On 7 July 2026, the EDPB adopted Guidelines 03/2026 version 1.0 for public consultation; the text was published on 8 July, with a feedback period until 30 October 2026. As of the date of this report, the document remains a draft recommendation, not the final version adopted following the consultation. This has implications for its legal formulation: the draft's conclusions are essential for understanding the regulators' position, but it is necessary to distinguish the GDPR itself from the EDPB's current interpretation and the Observatoire ARGA's recommendations [2][13].

The scope of the project is deliberately limited to scraping external internet sources by private organizations for the development of generative AI. The EDPB addresses cases of independent collection and procurement of previously collected data sets for training, but specifically does not fully cover the activities of brokers who simply offer data sets to others, as well as the collection of their own data by organizations and public authorities. This limitation of the document's scope does not exempt the excluded situations from the GDPR [2].

EDPB Opinion 28/2024 contains a relevant general position on personal data in the context of AI models, including the conditions for possible anonymization, legitimate interests, and the impact of unlawfully processed data on the subsequent use of the model. Therefore, the study uses Guidelines

03/2026 as a specialized development of the approach, and Opinion 28/2024 as additional material for the analysis of subsequent stages. The actual status of each version is verified separately before public use [3].

5. When does the GDPR apply to public internet data?

The GDPR applies to the processing of personal data, not to abstract internet documents as such. Identity is assessed based on the ability to directly or indirectly identify a person, taking into account the means that could reasonably be used, the context, and the ability to link attributes. A name is not a prerequisite: a persistent pseudonym, a photograph, a profile URL, a rare set of biographical circumstances, or a geographical feature may lead to identification [1].

Mixed collections of technical text, images, and personal messages do not become completely "impersonal" due to the predominance of publicly available articles. If the collection contains identifiable individuals, GDPR rules apply to the corresponding processing. For independent verification, it is useful to distinguish units of assessment at the source, document, and fragment levels, but the risk of identification when subsequently combining components must be kept in mind [1][2].

The exception for household or personal activities should not automatically extend to industrial crawlers. A legal entity creating a commercial or research generative model must first establish its own role in processing and the territorial applicability of the regime, then determine the legal basis and protective measures. Publication of information by the subject itself also does not constitute consent for any reuse purposes.

In a dispute, evidentiary weight is not only given to the user's visible web interface, but also to technical circumstances: archived copies, registration restrictions, information about who published the content, the date of the request, access control, the site's terms, and the possible removal of the material from the original source. The absence of such records makes it difficult to establish the context of legality and the user's reasonable expectations.

Within the GDPR, it's important to distinguish between information that constitutes personal data and images or text where identifying an individual is not actually possible by reasonably probable means. Legal classification is not determined by file format. A document without a clear name may contain a rare combination of job title, location, date, and circumstances sufficient to identify an individual. Conversely, general statistical indicators that do not allow for the identification of a specific individual may fall outside the scope of personal data, depending on the circumstances. Therefore, for random checking, the feasibility of combining information should be assessed, not just the availability of regular expressions for recognizing names and phone numbers. The risk of identifying an individual increases with vectorization and linking multiple indexes, although embedding itself does not necessarily constitute personal data in all cases.

The territorial scope of the GDPR requires a separate analysis under Article 3. The mere presence of a European domain name, the French language, or a European citizen on a website does not constitute a universal criterion for applicability. What is important is the activity of the controller or processor, and for entities outside the EU, the conditions for offering goods or services to entities within the Union or monitoring their behavior within the legally established limits. Separate consideration must also be given to who actually determines the purposes and means of processing, how the corporate groups are structured, and where the data is accessed. It is useful to document the

basis for the territorial connection and the degree of certainty in the report. This prevents both the unlawful extension of the GDPR to all global internet data and the equally erroneous assumption that the offshore location of a contractor automatically excludes European obligations.

6. Controller, processor, joint controller and seller of data

The status of a participant is determined by the actual decision on the purposes and essential means of processing, and not merely by contractual inscription. The developer, who defines the subject categories and requirements for data and instructs the contractor to technically collect them according to documented instructions, often acts as a controller, while the contractor may be a processor; when independently determining the purpose and set of sources, the roles may differ [2] [4].

If a broker first collects a dataset for its own purposes and then sells it to a developer, the initial collection and subsequent training must be assessed as separate operations of two independent controllers. A contract for the sale of a "licensed dataset" does not replace an analysis of the conditions under which the personal data was obtained. If the parties jointly determined the purposes and significant means, a joint controller regime may arise under Article 26 of the GDPR [1][2][4].

The role matrix must take into account the cloud platform, the labeling contractor, the cleaning studio, the embeddings manufacturer, and the downstream integrator, if each has access to personal data. For processing on behalf of the data controller, the following conditions are relevant: subject matter and duration, instructions, security, assistance to the controller, deletion or return of data, audit, and subprocessors. Affiliation between companies does not eliminate the need for such an analysis [1][4].

An independent auditor primarily verifies the actual data flow. A risk indicator is when a company claims that "the supplier is already responsible for GDPR compliance" but fails to identify the controller of the initial extraction, specify the purpose of the transfer, or disclose the volume of data being audited. Such a discrepancy is recorded as a lack of evidence, not automatically as a proven violation: a legal conclusion requires an examination of the facts.

A typical four-party chain illustrates why the contract title doesn't define the legal status: the platform publishes the material, the aggregator technically extracts the data, the broker normalizes and sells the dataset, and the developer selects the training and quality criteria. For each operation, it should be established who decided which categories to collect, on what basis, for how long to store, with which systems to combine, and to whom to disclose. The aggregator may be a processor on assignment, an independent controller in another part of its work, or a participant in joint goal-setting. It is necessary to examine the actual powers and decision-making methods, rather than formally assigning all contractors the status of "processor." When combining roles, the matrix should describe individual processes and documents, including the obligations of Article 28 of the GDPR and the requirements of Article 26 for joint controllers where they are truly applicable.

When purchasing a prefabricated corpus, the buyer does not automatically obtain a legal basis from the supplier's contract. It is necessary to check whether the original processing was lawful, whether the new purpose meets the compatibility criterion or requires a separate basis, whether the conditions of transfer have been met, and whether the rights of subjects can be exercised. The broker's guarantee of "full legality" is useful for allocating civil-law risks between the parties, but

does not in itself replace public-law obligations. It is also necessary to separate personal data from rights to protected content: the presence of a license for texts does not resolve the issue of Articles 6, 9, and 10 of the GDPR, and the presence of a GDPR basis does not preclude copyright verification. If the supplier refuses to provide even general information about the origin and prohibited categories, the insurmountable uncertainty should be documented and the admission of high-risk collections should be suspended, rather than simply raising the contractual risk assessment.

7. Availability of the page and the legality of its further use

Public access to the page affects individuals' legitimate expectations and the technical feasibility of obtaining it, but does not constitute an independent legal basis under Article 6 of the GDPR. If data may be processed for legitimate interests, it remains necessary to fulfill the entire proportionality test, assess the scope, types of information, the existence of objections, and minimization measures. Additional restrictions apply for certain types of information [1][2][5].

In terms of potential damage, an institution's publication of its schedule, an employee's general professional profile, a personal medical history, and a child's photograph are fundamentally different. All four examples may be accessible externally, but the individual's reasonable expectations, the purpose of the initial publication, the sensitivity, the permissible volume, and the risk of further dissemination are not aligned. The blanket rule that "everything publicly available may be collected" is legally untenable.

When processing content, it is necessary to evaluate the resource's contractual restrictions, applicable law regarding technical access measures, database protection, and intellectual property rights. Robots.txt, CAPTCHA, and other website objections are considered important factual and organizational signals in the legality and fairness analysis. They are not a universal substitute for a GDPR assessment, but they cannot be ignored without justification [2][6].

The decision-making matrix should include a separate "blocking" state for situations where the legal status of a sensitive set is unknown, not just three risk-level gradations. In such cases, temporarily suspending collection pending verification is a justified management outcome, not a declaration of the entire model's illegality.

8. Limiting the Goals: A Research Formula Instead of the "AI Training" Slogan

Article 5(1)(b) of the GDPR requires specific, explicit, and legitimate purposes. The term "improving artificial intelligence" is too broad to allow for a practical assessment of the necessity and scope of processing. The developer must specify the purpose of the model, its linguistic, sectoral, and functional scope, the intended users, the success criteria, and the reason for the need for the specific classes of personal data in question [1][3].

Training a common language component, configuring a medical assistant, training a human identification system, and developing a court document search engine all have different goals. The same initial web corpus cannot necessarily be reused for each of these purposes. When transitioning to a new goal, its compatibility, connection to the original publication, the nature of the data, and the implications are assessed; "the developer decided so later" is not a valid reason.

The meta-goal of "verifying security" doesn't make it safe to retain all identifiers. Quality assessment and incident response kits must also have a clear scope, duration, access, data type, and deletion process. The more uncertain the lifecycle, the more difficult it is to convincingly demonstrate minimization and compliance with the original legitimate interests.

It is recommended to approve the objective at two levels: the legal processing card (why and on what basis) and the technical specification (which documents, features, formats, tools, and metrics). Changes to both versions are approved before re-collection, not retroactively. The auditor verifies the versions and actual logs, not just the latest, prettier policy version.

9. Legitimate interest: the criterion of certainty and legitimacy of the goal

When relying on Article 6(1)(f) of the GDPR, three cumulative conditions must be demonstrated: the legitimate interest of the controller or a third party, the necessity of the processing to achieve that interest, and the absence of an overriding interest or fundamental rights and freedoms of the individual. An interest cannot be declared legitimate simply because the activity is innovative, commercially significant, or in the public interest. The EDPB emphasizes the need for a specific purpose and evidence, rather than an abstract promise of future benefit [1][2][3][5].

The first stage of the LIA documents who exactly benefits and how it relates to legitimate activity, what functional results are expected, and what legal restrictions apply in parallel. Commercial interest may be legitimate in nature, but in itself is not sufficient to conclude on the permissibility of processing. The document must allow an independent reviewer to critique the stated purpose based on observable facts.

If the stated interest is in ensuring the quality of legal translations, it should be explained why personalized online court decisions are needed for training, rather than an anonymized professional corpus. If the benefit is described as fraud prevention, it is necessary to clearly demonstrate the connection to specific fraud detection tasks, the potential use of sensitive information, and the limitations of subsequent purposes. These issues cannot be replaced by reference to the interest in abstract security.

In the absence of evidence of legitimate interest, another permissible basis should be sought or the processing should be abandoned. The report deliberately avoids asserting that legitimate interest is always the only possible basis for any AI-related processing: the basis is determined by the operation, its specific facts, and the limits of the relevant law. For mass private web scraping, the ICO considers legitimate interest to be practically the primary candidate in the typical circumstances it describes [7].

To formulate a legitimate interest, a verifiable connection between the business objective and the specific use of the data is required. The formula "development of modern technologies" is too broad to justify collecting all pages and storing them indefinitely. A narrower goal could be improving the quality of terminology translations for specific language pairs for a service available to the user. However, even such a goal alone does not justify collecting identifying information and does not prejudge the balance. The LIA should describe who will benefit from the intended use, why the goal is legitimate, the time horizon over which it can be achieved, how the expected usefulness is supported, and how the initial assumptions will be revised. If the model's purpose, its area of application, or its intended audience changes, the existing formula of interest must be revised.

The control conclusion must be falsifiable: can it be shown that the organization rejected a particular class of data precisely because it is not needed for a legitimate purpose? If all proposed classes are automatically deemed necessary, the document resembles a justification for the decision rather than an independent assessment. Records of rejected sources and a written justification for alternatives, including the use of licensed corpora, synthetic materials, aggregated data, and more narrowly targeted scraping, are helpful. The internal committee may require comparable model quality results without the most risky segment before authorizing additional collection. This is a methodological suggestion, not a specific, statutory procedure; the legal outcome always depends on the factual test of Article 6(1)(f) and the principles of Article 5 of the GDPR.

10. Legitimate interest: the criterion of necessity and alternatives

Necessity is not the same as convenience, affordability, or the advantage of more data. The second stage of LIA compares reasonable means of achieving the stated goal that are less intrusive on privacy. These may include licensed corpora with verified provenance, de-identified materials, selective subsets, high-quality synthetic data, localized cleaning, or limited collection based on pre-defined topics [2][6][7].

It is technically possible to reduce the volume of retrieved pages, the number of identifiers, the historical depth, the update frequency, and the storage time. It is necessary to specify which specific quality metric the restriction is expected to worsen and to what extent this is supported by testing, rather than intuition. A comparative approach to the stages of model development and use is also discussed by the ICO [9]. If material is collected "just in case" for uncertain future tasks, the necessity test is especially difficult to pass.

In its draft Guidelines 03/2026, the EDPB cites narrowing the collection criteria, pseudonymization, and, in some cases, synthetic data as means of reducing intrusiveness. However, pseudonymization does not automatically remove a set of data from the scope of the GDPR. Technical alternatives must be tested for leakage, realism, and the preservation of hidden identifiability; data that is merely renamed using hashes may remain personal [1][2].

Suggested evidentiary standard: For each rejected alternative, provide its description, expected functional loss, empirical support, additional burden, and the result of a comparative rights impact assessment. A blank table indicating "no alternatives" is not considered a reasoned answer. Measurability of such a choice is particularly important for independent audits of large collections.

Necessity is assessed based on the chosen legitimate purpose, not by comparing the processed data with the maximum technical capabilities of the infrastructure. Limitations apply not only to the number of records. The depth of the historical archive, repeated extraction, identifying metadata, the preservation of full-text documents when fragments are sufficient, and the use of biographical attributes in tasks where they are not needed are also important. For example, when training legal terminology recognition, it is permissible to test whether the research value is retained by excluding the names of parties, health information, addresses, photographs, and direct quotes from private publications. If the goal can be achieved in a comparable, less intrusive manner, the rejection of this method requires compelling justification.

The comparative alternatives method must consider not only the final test scores but also costs, data quality, practical applicability, and the impact on affected parties. Zero risk cannot be claimed in the abstract, but commercial convenience does not equate to legal necessity. The experimental results

are presented as limited technical evidence: a description of the A and B corpora, model parameters, fixed test tasks, result dispersion, and decision criteria. Testing on a small random sample does not provide grounds for declaring an alternative unsuitable for all languages and all tasks. Internal records will allow for subsequent verification that the declaration of "no alternative" was not used merely to justify cheaper, indiscriminate hijacking.

11. Legitimate interest: balance of consequences and expectations of the subject

The third stage involves a genuine comparison of legitimate interests and the interests, rights, and freedoms of individuals. The analysis includes the source, the context of the original publication, the predictability of repurposing, the breadth of subsequent use, the scale of processing, the level of observability, and the possibility of objections. The presence of an open profile does not imply the expectation that the content will be fed into a model capable of generating similar text or constructing individual profiles [1][2][3].

The risk profile increases when collecting data on children, private lives, political or religious interests, financial status, location, accusations, and materials older than several years, the content of which may no longer be relevant. The "chilling effect" on freedom of expression is also significant: awareness of total machine archiving can change people's behavior even without revealing an individual's name [2].

Mitigating measures reduce the likelihood or severity of harm, but do not serve as a blanket absolution. Domain restrictions, expedited takedowns, accessible collection notifications, objection mechanisms, separate filtering of complex categories, and memorization testing are considered only to the extent that they are actually in effect and verified. The formal existence of a policy does not equate to its implementation.

The interest statement model requires two independent signatures: the responsible business unit confirms the need, and the data protection officer analyzes the rights and remaining risk. If a material conflict remains unresolved, escalation and suspension of data collection are recommended if permissibility of processing cannot be ensured. This is a proposal of the ARGA methodology, not a standalone mandatory "dual-signature" mechanism under the GDPR.

The balance of interests must take into account the context of the initial disclosure of the material. Posting a message in a professional directory to connect with clients, sharing experiences in a self-help group, posting a court decision in a public database, and releasing a child's photograph by a parent differ in purpose and expectations. A user may reasonably expect other visitors to read the post, but they cannot expect to permanently retrieve all publications for commercial training and subsequent reproduction of details years later. Therefore, the formal openness of a resource does not replace an assessment of the scale of collection, audience, additional purposes, and potential adverse consequences. Information about health, political views, vulnerable individuals, and wrongfully disseminated accusations carries particular weight.

When assessing risk mitigation measures, it's important to check not only their presence but also their residual effect. A hypothetical scenario: a filter removes direct phone numbers from 99.5 percent of tagged test messages, but allows for rare combinations of last name and place of work that could identify a person. The first result doesn't prove overall security. Different data types require different criteria and Type II error checking. A positive LIA decision should include the

following conditions: precise sources, test update period, review period, complaint procedure, responsible person, and implementation oversight. When changing the model from an internal experimental tool to a widely available product, the balance may shift significantly due to the new scale of adoption.

12. Other grounds for processing, consent and false universal solutions

The GDPR provides six grounds for Article 6(1). Their applicability cannot be selected based on the convenience of reporting: consent requires compliance with the conditions of Article 7 and the possibility of its withdrawal; contractual grounds are related to the need to perform a contract with the data subject; and legal obligation and public purpose have independent legal prerequisites. For private mass downloading of other people's websites, these grounds often do not correspond to the actual circumstances; however, a categorical ban on their use in all models would be inaccurate [1] [7].

The collection of materials specifically provided by participants with informed consent is fundamentally different from the extraction of third-party publications that have not interacted with the developer at all. An author's consent to publish a website cannot be interpreted as the consent of all individuals mentioned in that publication to the subsequent mass training of a generative model. The legal basis is assessed for each operation and the relevant categories of personal data.

Withdrawal of consent, objection to processing based on legitimate interest, and prohibition of certain subsequent purposes have different legal natures. Therefore, a single "unsubscribe from AI" button is useful as an interface, but the internal processor is obliged to classify the requirements under the relevant articles of the GDPR. It cannot promise unconditional removal of all information from the model without analysis if the actual technical feasibility and legal obligation differ [1][10].

When combining different legal grounds, the distinction between sets and stages must be documented. It is unacceptable to "switch" the grounds retroactively as a way to justify a completed collection after an objection has been raised. Changes in legal position must be recorded openly, with an explanation of the facts and the consequences for individuals.

13. Special categories of data: Article 9 of the GDPR

Data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, as well as data concerning health, sexual life or orientation, genetic data and biometric data for unique identification fall under the special protection of Article 9 of the GDPR. Processing is generally prohibited unless a specific exception is provided for in Article 9(2) or an independent ground under Article 6 exists. The reference to the "public internet" does not compensate for the missing exception [1][2].

Web scraping is characterized by the problem of incidental acquisition: the developer searches for text of a specific genre, but also obtains information about the author's health, legal status, or beliefs. EDPB Draft 03/2026 considers the limited application of the logic of GC and Others C-136/17 to such incidental residual collection in the presence of effective preventive measures, while emphasizing the absence of a general exemption from Article 9. Whether such an argument can be relied upon is determined on a case-by-case basis, rather than automatically [2][17].

A practical procedure should distinguish between preliminary filtering of sources, filtering of already obtained material, temporary quarantine storage, removal from the dataset, and

investigation of filter bypass cases. An open dataset cannot be classified as "safe" solely because the search string did not include sensitive words. The semantic meaning of the fragment, the context, and the possible model inferences are also important.

If an increased risk is identified, collection is suspended for the affected category until the legal prerequisites are verified. The auditor requests frequencies and examples of false negatives for the filter, not just the stated detection rate. The verification result is formulated with an indication of the test sensitivity and residual uncertainty; zero error in a small sample does not equal the absence of risk.

The practical distinction between random and systematic collection is particularly important for Article 9. An independent auditor must determine whether the developer uses categories of websites that are highly likely to contain information about health, sexuality, or religious beliefs, and whether these are excluded prior to collection. Passing records through a filter after they are fully stored is not always equivalent to preventing processing: the collection and initial storage themselves already constitute data manipulation. Therefore, the audit includes domain exclusion logs, early filtering architecture, methods for restricting access to quarantined materials, and criteria for destroying residual copies. The assertion that "we did not intend to collect such information" requires evidence of action, not just a statement of intent.

The EDPB Guidelines 03/2026 draft discusses the applicability of the CJEU's reasoning in GC and Others to specific situations of incidental and residual collection, but it does not provide a blanket license for the processing of special categories. The analysis remains case-specific and takes into account the controller's actual ability to prevent collection and further dissemination. For content clearly falling within such categories, an independent assessment of the applicable conditions of Article 9(2) is required, along with the basis of Article 6. In the absence of a processing condition and the impossibility of reliably excluding the relevant segment, the methodological outcome may be a complete rejection of a certain class of sources. This legal dilemma cannot be replaced by a general promise that the model "does not remember sensitive data anyway."

14. Information on convictions and offenses: separate article 10

Article 10 of the GDPR establishes a special regime for the processing of data relating to criminal convictions and offences or related security measures. Such processing, pursuant to Article 6(1), shall be carried out under the supervision of an official body or in cases permitted by EU or Member State law with appropriate safeguards. The publicity of the judicial document or the mention of the criminal case does not derogate from these requirements [1].

When monitoring international wanted lists and media mentions, the risk is the automatic conversion of charges, outdated publications, acquittals, and name matches into machine-generated indicators of "connection to a crime." Research must strictly distinguish between the court-established legal classification, the stage of the investigation, press reports, and mistaken identity. The generative model is not an official source of information about the existence of an international notice or verdict.

If a training corpus contains such information, the developer must demonstrate the specific legal justification, processing conditions, and documented safeguards. The general statement regarding the open nature of court registers does not replace national restrictions on access, reuse, and retention periods. The use of personal data to evaluate a specific individual requires independent

verification of the legal framework, including Articles 5, 6, 10, and 22 of the GDPR, where applicable [1].

Recommended audit controls: a selection of legal and news sources with markers for identity accuracy, event date, change in procedural status, and official status. In a disputed situation, the recording is assigned the status "not verified" rather than a negative characterization of the individual. Independent verification should prioritize protecting the individual from the automated dissemination of false incriminating information.

Information about criminal convictions and court cases constitutes a separate risk group. Article 10 of the GDPR is not a variant of Article 9 and establishes special requirements for the processing of personal data about criminal convictions and offences or related security measures. The assessment includes the national legal basis, the existence of official oversight in the specified cases, and the scope of permissible processing. It cannot be assumed that the public publication of an indictment, a notice of inquiry, or a copy of a court document automatically permits the commercial extraction of such information for a generative model. Furthermore, the standard filter for disease names does not address the issue of criminal procedural information, which is often contained in free text.

A practical mechanism must distinguish between suspicion, indictment, conviction, dismissal of a case, dismissal of charges, and refusal of extradition. These events have different legal and factual meanings, and mixing them can cause serious reputational and financial damage. An independent review examines the source documents, the date of the event, the subsequent status of the proceedings, and any indication that the decision is not final. Background or journalistic reports should not be turned into a "universal criminal profile." If processing restrictions are in place, even accurate information may be inadmissible for a specific purpose. The report must clearly indicate whether the conclusion relates to the quality of the model, the legality of the original collection, the legality of storage, or the subsequent use of the result by a financial institution or government agency: these issues are subject to different legal tests.

15. Data of minors and vulnerable persons

Child protection is particularly important when balancing Article 6(1)(f) of the GDPR: the rights and freedoms of individuals, particularly children, may outweigh legitimate interests. School publications, photographs, adolescent forum posts, and family pages may be included in training materials. The technical publicity of such materials does not ensure the predictability of their use in the generative model [1][2].

The developer must define prohibited categories of sources, modes for detecting minors, rules for removal in case of doubt, restrictions on the use of emotionally sensitive materials, and accessible methods for legal representatives to contact. Age-detection errors can affect both adults and minors; therefore, the protocol must account for false positives and provide for secure verification without the need for additional document collection.

The ARGA methodology recommends against collecting specific age-related confirmation information from third-party registries simply to "clarify" the compiled corpus if this would increase rights infringement. It's better to use restrictions based on domain type and context, minimize collection rules, remove the disputed class, and provide a clear reporting channel. Each source exclusion is documented, with a reason and the possibility of reconsideration.

If data has already been used for training, mitigation requires assessing the model's behavior, the ability to extract content, and methods for blocking the output. It's essential to honestly distinguish between the technically proven effect of measures and the intended effect. Promising "total forgetting" without a verifiable protocol can further damage trust and violate transparency requirements.

16. Transparency requirement and Article 14 of the GDPR when data is obtained indirectly

When personal data is not collected from the data subject, Article 14 of the GDPR requires disclosure of the controller, purposes, grounds, categories of data, source, and other circumstances of the processing. In the case of mass web scraping, individual notification may be difficult in practice. However, this difficulty is subject to justification under a specific regulation, and is not a basis for eliminating transparency [1][2][14].

EDPB Draft Regulation 03/2026 specifically addresses Article 14(5)(b) of the GDPR when providing information proves impossible or requires disproportionate effort. Even if the conditions for an exception are met, the controller must take appropriate measures to protect rights, and publicly posting the necessary information is a legally required measure. Silently taking advantage of the exception and not publishing anything is not permitted [2].

A clearly understood public policy should describe the types of sources and download periods, categories of personal data, purposes, specific legitimate interests, retention periods, risks and rights of data subjects, objection channels, an exception handling mechanism, and controller contact information. A searchable source register or, at a minimum, a structured list of domains and resource types with an explanation of the incompleteness of the list if publication of each URL is not possible is recommended [2].

Transparency must begin before mass collection and be maintained as the model's goals, sources, and architecture change. Text buried in hundreds of pages of specifications and lacking a practical search mechanism may formally exist, but it won't provide effective information. To verify this, it's important to conduct a user test: is a stakeholder capable of understanding the situation and initiating a request without legal or technical knowledge?

When data is obtained indirectly, the controller must provide the information required by Article 14 of the GDPR, including the data categories, purposes, legal basis, relevant sources, recipients, and the retention period or criteria for determining it. A simple reference to the general "use of AI" does not allow individuals to understand how their photo, publication, or archived recording could have ended up in a training corpus. A two-tiered notification structure is helpful: a short, clear description with a rights channel and an expanded technical section with website categories, data collection versions, and types of operations. However, even a detailed public page does not automatically serve as a full substitute for individual notification in all circumstances. The applicability of exceptions is assessed separately, and the method of publication must ensure reasonable accessibility of the information to affected individuals.

For international systems, it's important to consider the languages and actual means of discovering a notice. Posting an English-language PDF in the technical section of a website is unlikely to provide meaningful transparency for people whose messages are collected in other language communities. One can evaluate how easy it is to find a notice in search results, whether it includes contact

information for exercising rights, whether it explains obscure technical terms, and whether it provides a channel for reporting a disputed source. This is primarily an assessment of the effectiveness of the proposed measures. Compliance with the specific obligations of Article 14 is determined by law and facts. For the Observatoire ARGA pilot program, a test with independent participants asking them to answer three practical questions would be appropriate: why my data may have been used, how to verify the circumstances, and how to request a correction.

The deadlines for provision of information when data is obtained indirectly are determined by Article 14(3) of the GDPR: within a reasonable time after receipt of the data, but no later than one month; if the information is used to communicate with a person, no later than the first communication to him or her; if disclosure to another recipient is intended, no later than the first such disclosure. Prior to further processing for another purpose, the separate requirement of Article 14(4) applies. The exceptions of Article 14(5), including impossibility or disproportionate effort, apply only in the presence of demonstrable conditions and do not constitute a general exemption for all large teaching corpora [1][14].

17. Disproportionate effort criterion: the exception must not swallow up the rule

When analyzing Article 14(5)(b), the actual burden on the controller must be weighed against the consequences of failure to notify the individual. The EDPB draft cites the following factors: the number of data subjects, the age of the data, and the existing safeguards. It specifically cautions against the automatic application of the exception outside of specific archiving purposes, scientific or historical research, and statistics. The proposed justification must be reproducible and take into account the dataset as a whole [2].

For example, tens of millions of old documents without direct identifiers can create an unusually high cost for individual notification. However, this doesn't necessarily mean that any commercial fee is always considered "disproportionate." If the control set contains contact information for a small group, notifying them may be technically feasible. A single legal memo for both cases obscures the relevant differences.

The decision log should record the dataset characteristics, an assessment of the communication methods, the impossibility and risks of reverse identification, the cost and quality of alternative information, the explanation for the exception, and any additional security measures. Any evidence of disproportionality should be linked to a specific version of the dataset and the time of the decision.

Observatoire ARGA recommends maintaining two interconnected documents: an internal reasoned decision under Article 14(5)(b) and published, understandable information for subjects. The internal report should not disclose sensitive information in the public version or allow for re-identification. If one part is inconsistent with the other, the matter is subject to further review.

18. Minimization: Before, During, and After Extraction

Article 5(1)(c) of the GDPR requires data to be adequate, relevant, and limited to the necessary volume. In the planning stage, this means excluding irrelevant domains, sections, and document types. During extraction, this means limiting fields, crawl depth, time period, access frequency, and identifiers. After collection, this means cleaning, deduplication, removal of unnecessary fields, and reasonable reduction of retention periods [1][2].

Systematic application of the minimization rule is more important than a one-time cleanup before training, as pre-collected data may be stored in intermediate storage, backups, and logs. If a prohibited set of data has been "cleaned" from the final file but remains in the raw bucket accessible to developers, the minimization goal has not been achieved. Audits should verify copies, access rights, and retention procedures.

The EDPB proposes to consider synthetic, pseudonymized, and anonymized data, as well as restricting websites that structurally contain special categories. However, synthetic material cannot be considered inherently risk-free: generation based on real samples can reproduce personal information. Acceptability criteria should include similarity and leakage tests and a confirmed procedure for the destruction of original copies [2][15].

An important management issue is the timing of filtration. Filtering before extraction generally better limits the actual processing; post-receipt cleansing remains a necessary safeguard against erroneous capture. Both lines of control complement each other, but do not replace the legal basis for the original operation.

Minimization at the design stage begins with a negative list. In addition to permitted website categories, it's important to pre-list sources that, by their very nature, often contain disproportionately sensitive information: self-help resources for the sick, children's communities, private address aggregators, forums for survivors of violence, archives of court charges, and websites that specifically object to automated collection. Specific criteria can regulate geography, publication age, URL structure, and the presence of direct identifiers. Regex filters should not be used alone during development: names, health information, and legal status are expressed contextually, and classifiers are prone to errors in rare languages and indirect wording. Therefore, filtering requires independent verification for false negatives.

After collection, a multi-level processing is proposed: automatic highlighting of potentially risky data, limited quarantine with enhanced access control, targeted verification, and deletion or depersonalization, if feasible. Each stage stores the counts of excluded records by category, the proportion of questionable cases, and the rule version. The "black box" problem arises when the contractor returns only the final number of deletions without the original calculation base and verification mechanism. The statistic "two million records deleted" cannot be used to determine how many risky documents remain. For review, an independent sample of unfiltered records, separate measurement by language and content classes, and periodic replay of the test after rule updates should be used.

19. Reliability, time and the problem of outdated information

The accuracy principle under Article 5(1)(d) of the GDPR requires that personal data be accurate and, where necessary, up-to-date. For generative learning, data from forums, news, and old registries may reflect a state of affairs that has since changed. A model that combines old text with a current question may produce an erroneous legal or biographical profile [1].

In its draft Guidelines 03/2026, the EDPB recommends selecting reliable sources, recording the time of extraction, and verifying data before using it for training. This is particularly important in sanctions lists, international investigations, commercial compliance, and criminal proceedings, where records may have been amended, deleted, or restricted in access [2].

Separate attributes should be stored: event time, publication time, download time, and last check time. A single "scraped_at" timestamp does not prove the relevance of the event being recorded. The verification procedure should distinguish between failure to update, conflicts between sources, and a confirmed status change. For high-impact events, a source with expertise in the specific legal fact is required.

In measuring quality, the error will be the proportion of texts "without typos" instead of the accuracy of disputed claims about individuals. The proposed metrics include: the number of sources with missing dates, the proportion of negative personal statements without primary confirmation, the latency of updating a changed record, and the proportion of restrictions on the use of outdated information actually applied to training samples.

20. Data protection impact assessment and risk documentation

Article 35 of the GDPR requires a DPIA when the type of processing is likely to result in a high risk to the rights and freedoms of individuals. Large-scale automated collection, innovative technologies, sensitive data, and non-transparent use may give rise to such indicators; however, the need for a DPIA is determined on the basis of specific facts and applicable supervisory authorities [1][16].

A proper DPIA describes the operations and purposes, necessity and proportionality, risks to individuals, and the intended mitigation measures. Of particular importance are risks that cannot be detected solely by unauthorized access: incorrect re-identification, reputational damage from generative inference, false sanctions matches, mass indirect profiling, failure to exercise objection, and hidden persistence of information in the model.

The assessment methodology should clearly distinguish between the likelihood of impact, the severity of harm, and the quality of available evidence. High uncertainty should not automatically reduce the level of risk. The threshold scales proposed in the appendix serve as an organizational tool; they are not legally established categories. If the residual high risk cannot be mitigated, conditions for prior consultation with the supervisory authority may arise under Article 36 [1].

The assessment is revised when objectives or source classes change, a new data broker is added, scale significantly increases, or the model is substantially modified. A retroactively rewritten document without a review history does not ensure accountability. The independent assessor must have access to the original assessment, a log of updates, and an explanation of why the controls were deemed sufficient.

The determination of whether a DPIA is necessary under Article 35 of the GDPR cannot be mechanically inferred from the sole criterion of "training artificial intelligence." An assessment of the nature, scale, context, and purposes of the processing is required, including new technologies, the potential impact of systematically assessing individuals, aggregation of data, vulnerable groups, sensitive information, and legal protections. For large-scale, non-transparent collection of personal data, the risk may be high, but the final conclusion must be based on the actual project design. The assessment document describes the proposed operations, their necessity and proportionality, the risks to rights and freedoms, safeguards, responsible parties, and the residual risk. A formal, color-coded matrix without supporting reasoning is insufficient.

If, after reasonable measures, a high risk remains that the controller is unable to mitigate, the issue of prior consultation with the supervisory authority is considered under Article 36 of the GDPR.

This procedure is not replaced by the statement "our DPO has approved the launch." An independent audit also cannot guarantee that the presence of a DPIA guarantees the lawfulness of any subsequent operations. For practical verifiability, it is proposed to link each significant risk to an observable test, evidence, and review condition. For example, the risk of facial recognition through a combination of biographical details is linked to testing for linking multiple sources; the risk of memorizing personal phrases is linked to a controlled extraction assessment. The frequency of reassessment depends on changes to the corpus, model, sources, and the nature of the service provided.

21. Right of access, rectification, erasure and restriction of processing

The rights of data subjects are enshrined in Articles 12–22 of the GDPR and depend on the terms of each right. When making a request regarding web scraping, the developer must determine what data about the individual is being processed, in which versions of the data sets it is stored, where corrections are available, and how the request relates to the model used. The lack of a common search engine across sources complicates the fulfillment of legal obligations, but does not eliminate them [1][10].

The right to erasure under Article 17 is not unconditional in all circumstances, just as an objection under Article 21 does not have the same consequences for all legal grounds. However, the controller is obliged to ensure a reasonable and timely review process for the request, to provide reasons for the applicable exception, and to document the technical measures taken. The accessible form of the request must distinguish between deletion from the original corpus, prohibition against re-extraction, and measures to reduce the model's output.

Implementing a correction requires careful consideration of context. A legally significant historical fact may exist in a court news story, followed by a subsequent correction. A model trained only on the original publication may continue to convey the old status. The compliance procedure should identify conflicting versions and provide reasonable limitations on risky conclusions, rather than automatically rewriting the archival document.

The audit metric for rights effectiveness is not limited to the number of closed requests. It should measure the completeness of the search for relevant copies, the legitimacy of refusals, the validity of the chosen measures, the verifiability of the technical result, and the accessibility of a human appeal. The subject should not be forced to prove the "exact URL included in the dataset" if the controller's architecture does not provide such traceability.

When fulfilling data subject requests, the first step is identifying the subject and the controller. Not every "remove me from AI" request requires the same technical solution. The organization determines whether the record exists in the original archive, was included in the cleaned set, is used in the retrieval index, was fine-tuned, is stored in the query log, and whether the model is capable of reproducing the data without external context. The response must distinguish between the confirmed existence of the data, the actions taken, and the limits of technical verification. If some information cannot be reliably retrieved, it is unacceptable to misinterpret uncertainty as proof of non-processing.

For practical protection of rights, it is important to prevent collateral disclosure of information through the request channel itself. An access request may come from a person with a matching name or from an outsider seeking to determine whether the data set contains information about a

specific individual. The identity verification method must be proportionate to the risk and not degenerate into excessive document collection. The review timeframes and response procedures stipulated by the GDPR must be maintained on the controller's side, and the complexity of working with model scales should not automatically be used as grounds for indefinite silence. For disputes over criminal and sanction matches, a review path by a human expert and instructions on what primary evidence can help restore the precise context should be provided.

Procedural deadline under Article 12(3) of the GDPR: Action on data subject requests must be communicated without undue delay and, as a rule, no later than one month after receipt of the request. If necessary, the deadline may be extended by a maximum of two months due to the complexity or number of requests, but the controller must communicate the extension and the reasons for it within the initial month. If a decision is made not to take action, the requirements of Article 12(4) regarding a reasoned response and safeguards apply; additional identity verification is possible within the limits set out in Article 12(6). These deadlines should be set out in the request register and contractual SLAs, without promising technically unverified deletion of data from the model scales [1].

22. Objections to processing and taking into account website restrictions

A data subject's objection to processing under Article 21 of the GDPR and restrictions on automated retrieval indicated by a website owner are legally distinct. The former is a right of an individual, while the latter may reflect expectations, contractual provisions, or content controls. Both types of information must be documented, but excluding a domain from crawlers does not guarantee that an individual request about a specific individual will be fulfilled [1][2].

The EDPB and CNIL pay attention to technical mechanisms by which websites express their desire not to be automatically collected, including robots.txt and CAPTCHA. Ignoring these mechanisms increases compliance risks; however, the presence of a robots.txt file alone does not resolve the issue of excluding text from a previously collected set. The provider must separately manage the upstream and downstream versions of the corpus [2][6].

The proposed architecture maintains a deny registry with a scope of URL, domain, content type, specific subject, document identifier, and version. Each entry contains the reason for the denial, the date it was issued, the execution status, and associated operations: do not reassemble, exclude from further training, examine existing copies, and track complaint. If it is impossible to match a request to a technical object, the decision is not marked as fully executed.

A well-founded objection requires a separate legal assessment and a reasoned response. An application cannot be automatically rejected based on the developer's broader public interest or the supplier's lack of a mechanism to erase the influence of data from scales. The impossibility of the technology and the absence of a legal obligation in a specific form must be established separately, not mixed together.

23. Model weights, memorization and the question of "forgetting"

Training translates input observations into model parameters; individual training samples may remain reproducible for certain queries. Technical risk depends on the size, frequency, rarity, and nature of the data, the training algorithm, and the attack method. Verbatim extraction testing, sensitive sequence testing, and generative inference assessment can detect privacy violations, but do not exhaust the possible methods for identifying information [3][10].

From a legal perspective, deleting an object from a raw dataset, a curated training set, a retrieval index, and backups are distinct operations. Changing weights through machine unlearning, retraining, or model updating requires separate confirmation of the effect. The "deletion request fulfilled" statement must indicate which levels have actually been processed, which remain at risk, and why [1].

Often, a less risky interim solution is to exclude controversial fragments from subsequent retraining and separately protect sensitive data from being released. However, a protective release filter is not equivalent to erasing personal data from all processing, and it can be circumvented or may degrade the quality of legitimate responses. The report recommends testing the residual risk, documenting the limitations of measures, and, if necessary, requiring architectural changes.

Experiments with canary sequences or other test identifiers are only permissible in a legally established test environment, to prevent the test itself from becoming a new source of unauthorized processing. When exploring someone else's model, permitted interfaces, proprietary synthetic data, and adherence to access conditions are preferred. The report must distinguish between a hypothesis about memorization and a proven fact of disclosure.

The technical requirement to "forget" data isn't limited to a single operation. In an architecture with retrieval-augmented generation, an index can change quite quickly if the problem actually originated from it. In fine-tuning systems, reverting to a previous checkpoint, retuning on a cleaned dataset, or limiting the output are possible. When pre-training a large model, the influence of individual samples on the weights is difficult to isolate, and machine unlearning methods require independent verification of effectiveness and unwanted side effects. Removal from the original corpus primarily affects subsequent cycles and does not automatically confirm changes to the previously released model. It's impossible to promise universal, complete removal from parameters without qualification.

A gradation of evidence is useful for an independent conclusion. Level A: physical removal of a known record from controlled storage is confirmed. Level B: its exclusion from future training builds is demonstrated. Level C: behavioral tests of a specific model version have been performed according to a pre-established protocol. Level D: There is research evidence of data impact reduction, but it is not technically feasible to prove zero residuals under all conditions. The formulation of the result should be appropriate for the level, including failed tests and search limits. It is especially important not to generalize the evaluation of a single available model version to all copies, mirrors, and variants outside the control of the organization being tested.

24. Anonymization, pseudonymization and re-identification

Pseudonymization reduces the immediate availability of identifiers, but in the presence of additional information and the ability to associate a record with an individual, it remains the processing of personal data. Anonymization requires an assessment of identifiability, taking into account objective context and reasonably probable means. Simply replacing a name with a hash, random number, or initial does not guarantee anonymity [1][15].

In July 2026, the EDPB published draft Guidelines 02/2026 on Anonymisation, which considers contextual and simplified approaches and three tests: record extraction, linkage, and person inference. Their results do not replace a comprehensive legal analysis; the draft's status as of the date of publication remains consultative. When evaluating a large training corpus, it is necessary to

distinguish between the state of the data for the immediate controller, the recipient, and a potential attacker [15].

The practical anonymization test includes searching for rare attribute combinations, connecting with external sources, assessing individual participation, and the ability to reconstruct identities based on model output. The results are recorded, along with a description of attack scenarios, external data versions, assumptions, and limitations. Comparison with a closed database of unauthorized specialists cannot be described as a real test.

Unlike a declaration that "data is anonymized," an evidentiary assessment requires disclosure of the threat model, access rights to keys, the method for deleting the original data, and the results of re-checking when new linking sources are discovered. With an uncertain data status, it is safer to continue applying GDPR safeguards until a reasoned conclusion is reached, without presenting this approach as a legal definition of anonymity.

25. Origin Catalog: Verifiable Evidence for Each Source

Data provenance represents the chain from the original URL or file version to a specific training set and model. The registry must preserve the resource identifier, domain category, access period, upload date, public status, applied processing basis, restrictions, and transformation sequence. Aggregated or limited information may be used for publication if disclosing direct links would create additional harm [1][2].

A cryptographic hash allows one to establish that the submitted copies are identical between checks, but does not confirm the legitimacy of the extraction or the authenticity of the primary source. An electronic log with signed timestamps helps to ensure the sequence of operations, but without independent verification, it can only reflect the operator's assertions. Therefore, selective comparisons with source systems, network request logs, and supplier contracts are necessary.

Each transformed fragment must inherit source constraints, including prohibiting reuse. When aggregating thousands of documents, internal identifiers allow for the exclusion of individual lines of origin without revealing unnecessary personal details. The data architecture must allow for the correction of misattributions and the recording of source conflicts, rather than forcing a single version.

An independent audit determines the minimum traceability required: from the selected model output or training example back to the manifest, cleanup operations, and source. The inability to trace is a limitation of the audit and subject to recommendations. It does not automatically provide grounds for concluding that all material was obtained illegitimately.

Reliable traceability requires a link between at least four identifiers: the source, a specific record, a dataset version, and a model version. When website links change, it's not necessary to maintain a permanent, full copy of the disputed text; depending on the purpose and security requirements, hashes, limited metadata, and a separate archive with controlled access may be used. Each method has its limitations: a cryptographic hash confirms the match of a specific byte sequence, but does not prove its legal authenticity or explain its origin. A supplier's signature verifies the signatory, but does not transform unproven guarantees into established fact. Therefore, the registry contains information about the acquisition procedure, the trusted source, and verified checks.

Special requirements apply to the chain of subsequent changes. Suppose the original document is included in package D1, deleting identifiers yields D2, deduplicating D3, and merging with another corpus yields D4. If only the final D4 number and the general phrase "cleaned" are retained, it is impossible to determine which filter affected the disputed record and which models used it. It is recommended to maintain a transformation log with the procedure code, time, configuration version, number of excluded objects, and the linking package identifier. For independent audit, limited access to only the necessary fragments is possible, ensuring provability without requiring the entire corpus to be submitted to an expert. The legality of storing artifacts is also subject to separate assessment.

26. International transfers and third country access

When accessing personal data from a third country, the applicability of Chapter V of the GDPR is determined by the actual structure of the transfer, the roles of the parties, and the legal context. Hosting a server in the EU does not preclude an international transfer if an authorized foreign contractor accesses the data; however, not every use of a foreign cloud brand in itself constitutes an unlawful transfer [1].

The controller documents the recipient countries, the legal mechanism for the transfer, subprocessors, applicable contractual provisions, security measures, and accessibility under foreign law. It is necessary to distinguish the legal basis for the processing itself under Article 6 from the mechanism for international transfer under Chapter V: compliance with one does not preclude the other.

For global corpora, machine learning replication, temporary working copies in cloud environments, operator query logging, and shared debugging are essential. Inventory should cover not only the main training bucket but also the orchestration pipeline, monitoring messages, and images from external tools. Each technical route is fact-checked.

An independent auditor has the right to request evidence of cross-border risk assessment and contractual liability, but should not substitute a general "low/high" rating for a competent judicial determination of the legality of an international transfer. When only supplier statements are available, this limitation of the research data should be clearly disclosed.

27. Dataset security and contractor responsibilities

Article 32 of the GDPR requires appropriate technical and organizational measures, taking into account the risks of processing. For web scraping, this includes segmentation of raw, cleaned, and training environments, role delineation, access logging, key management, export controls, encryption where appropriate, and backup retention limits. Having a single "secure" cloud does not ensure the security of the entire data collection [1].

Website access tokens, cookies, accidentally collected confidential documents, outdated backup copies, and annotation subsystems with contractor access pose a distinct threat. It cannot be assumed that training on public data makes all resulting fragments suitable for unrestricted internal distribution. These risks require access testing at every critical stage.

The agreement with the processor under Article 28 of the GDPR must effectively provide assistance in exercising data subjects' rights, restrictions on subcontracting, erasure procedures, and access to evidence. A separate register of subprocessors, including changes to storage regions and access

categories, is useful. In the event of a security breach, the applicability of the obligations under Articles 33–34 is determined separately from the investigation procedure for illegal scraping.

An independent security assessment should include verification of the validity of audit trails, not just the presence of certificates. External certifications are acceptable as a measure of assurance, but their scope must be verified: an information security management certificate alone does not confirm the legality of processing each URL.

28. Intellectual property: a parallel, not an alternative, legal regime

Scraping may affect copyrighted works, the rights of database producers, and contractual access regimes. In the EU, Article 4 of Directive (EU) 2019/790 provides for an exception or limitation for text and data mining of legally accessible works, provided that the copyright holder has not properly reserved the relevant use. This exception applies within the limits of its terms and national implementation [18].

This does not constitute an exemption from the GDPR, as personal data and copyright protect different objects and interests. A work may not contain personal data but still be protected by copyright; a personal communication may lack significant creative originality but still fall under the GDPR. A collection may simultaneously incorporate both regimes, as well as provisions on commercial secrets and technical access protection.

For an independent audit, it is recommended to separate two sets of evidence: the legality of personal data processing and the permissibility of using the content for TDM/training. A lack of results on one side is not compensated for by a positive conclusion on the other. If rights reservations are detected, the detection technology, signal type, date, and impact on the actual versions of the corpus being used should be documented.

A public listing of training data under the AI Act cannot be interpreted as publishing all protected documents or as a legal license for them. The requirement for sufficient detail in the listing is intended to ensure transparency, not to convey the right to reproduce the entire source corpus. The specific terms of any exceptions or licenses concluded are subject to independent assessment [11][12][18].

29. AI Act: GPAI Documentation, Copyright, and Public Summary

Article 53 of the AI Act requires providers of general-purpose AI models to create and update technical documentation, provide certain information to downstream providers, have a policy for compliance with EU copyright and related rights law, and publish a sufficiently detailed summary of the content used for training, in accordance with the AI Office template. The rules apply according to the current calendar and transitional provisions; special exceptions for certain open models and separate obligations for models with systemic risk must be taken into account [11][12].

For web scraping purposes, it is legally significant that technical documentation under the AI Act and the GDPR processing records have different purposes, addresses, and compositions. The obligation to publish a summary of training content does not imply an obligation to publicly disclose every personal profile or protected contract; nor does it replace a verifiable internal register of the origin and grounds for processing personal data.

The report proposes a mapping: source category in AI Act summary ↔ internal dataset manifest ↔ legal basis/retention record ↔ license or rights reservation ↔ filtering protocol ↔ leakage test

report. This mapping avoids situations where a public list of categories appears convincing but has no connection to real models and loading histories.

The bodies authorized to implement the AI Act and the data protection supervisory authorities perform different functions. AI Act compliance reviews do not replace the supervisory procedures established by the GDPR. Therefore, recommendations for a single "lawful scraping certificate" should not be interpreted as existing state accreditation or universal authorization.

The timing and exceptions require separate verification. The obligations of Chapter V for relevant GPAI model providers began to apply on 2 August 2025. For models placed on the market before this date, Article 111(3) sets the transition date until 2 August 2027. Article 53(2) exempts some models with a genuinely free and open license and publicly available parameters from the obligations under 53(1)(a) and (b), but not from the copyright compliance policy and public summary of the training content under (c) and (d); this exemption does not apply to models with systemic risk. The GPAI obligations should not be confused with the timing of application of separate rules for high-risk AI systems. Establishing compliance with the AI Act does not in itself prove the lawfulness of processing each source record under the GDPR [11][12][19].

The Observatoire ARGA report uses a "double matrix of evidence": it analyzes the AI Act obligations separately and the GDPR separately. The AI Act column assesses the availability of documentation on the origin and content of training data, copyright policy, and public disclosure. The GDPR column identifies the objectives, roles of controllers, the basis of Article 6, the additional provisions of Articles 9/10, information, minimization, and the rights of subjects. The overlap can provide useful evidence, but does not replace separate tests. For example, the description of the source "public news sites" helps understand the class of content, but does not indicate whether unlawful archival allegations have been excluded or how the enforcement agent responds to the correction of the court record.

30. The French approach: CNIL and safeguards for online collection

The CNIL published guidelines for the development of AI systems and specifically addressed the collection of publicly available data through web scraping, including legitimate interest and mitigation measures. The French regulator emphasizes that web scraping itself is not prohibited by the GDPR as a technological operation, but its legality depends on the specific purpose, processing conditions, and safeguards [6][8].

For a practical project in France, it is important to document the vulnerability of sources, the use of exclusion lists, public disclosure, and the filtering of information not required for training purposes. Signs of website resistance to automated collection, including technical ones, cannot be blindly ignored. At the same time, the legal content of the CNIL recommendation should be distinguished from the express text of the GDPR, especially when national practice develops general provisions [6].

Evidence must be in plain language, describe the responsible controller, and describe the actual means of exercising rights. A policy on a French website that only lists corporate interests and does not describe the sources of personal data does not guarantee compliance with the principle of transparency. The auditor verifies the availability and relevance of information, not just the presence of a website.

The ARGA methodology proposes using the CNIL approach as a minimum check for transactions involving France, but not automatically transposing it as mandatory law for all third countries. For international cross-border projects, the applicable jurisdictions and substantive competence of each regulator must be determined.

31. UK: ICO position and comparative restrictions

The UK has its own UK GDPR and data protection regulations; the ICO's position is useful as a comparative reference but does not constitute a normative interpretation of the EU GDPR for all member states. In its response to the consultation series on generative AI, the ICO points out that legitimate interests in typical current mass web scraping practices are only realistically justified if they satisfy a three-part test, specifically necessity and a fair balance [7].

The ICO emphasizes the risk of opaque, extensive, and often invisible processing. If a less intrusive means of achieving the same goal is available, the controller must explain why it cannot be used. Similarly, significant deficiencies in transparency that prevent individuals from exercising their rights make the balancing of interests particularly challenging [7].

For comparative studies, it's important not to substitute a universal "UK permits - EU prohibits" chart for individual assessments of national laws. Real differences may lie in the industry-specific methods of supervisory authorities, procedural instruments, judicial practice, and the details of the regulated process. It's recommended to apply a map of regulations, effective dates, and competent authorities to each case.

An international group using the same dataset for models in the EU and the UK must conduct assessments at the transaction level, taking into account different geographic boundaries and contractual arrangements. A single global document may contain compatible general measures, but the burden of proof of applicability and compliance with the relevant regime remains.

32. Opaque kit suppliers: due diligence before purchasing

Purchasing a "ready-made, cleaned corpus" may postpone the technical work, but does not automatically provide the buyer with confirmation of the legality of the original collection. Verification must determine who collected the data, for what purpose, on what basis, what source restrictions were taken into account, whether special categories or criminal records are included, how the rights of subjects are protected, and whether independent verification is permitted [1][2][4].

Particular care is required when purchasing materials collected through a chain of brokers. If the primary supplier fails to maintain the source date and restrictions, subsequent guarantees easily become purely contractual. It is recommended to require a provenance manifest at the category level and representative samples with secure verification capabilities, as well as confirmation of exclusion of materials based on complaints.

The impossibility of disclosing the entire corpus for commercial reasons does not eliminate the need for sufficient evidence. Confidential data rooms, independent experts, sample audits without copying sensitive materials, and certification of specific procedures may be used. However, the auditor must clearly distinguish between partial access and complete confirmation of the absence of violations.

The contract includes the right to targeted audits, the definition of acceptable sources, notification of significant changes and incidents, deadlines for fulfilling legitimate requests, a procedure for

destroying copies, and the allocation of costs for corrections. All contractual guarantees are compared with the actual technical ability to fulfill them.

The procurement process for the training corpus should begin before signing the license agreement. The buyer develops a risk-based list of questions about the origin, timing, frequency of updates, previous collection purposes, site types, possible special categories, exceptions, and subject rights. If the broker assures everyone of "full anonymization," it is necessary to find out which identifiers have been removed, how the possibility of re-identification was assessed, whether linkage tests were conducted, and who has access to the original versions. The conditions under which the transferred records may effectively become personalized again after integration with other systems are separately examined. The inability to disclose commercially sensitive methodology does not always preclude a more limited independent verification in a secure environment.

Contractual provisions have practical value when they are linked to actions: the right to selective audits, the obligation to report newly discovered categories, prohibition of use of certain segments, notification of source changes, a mechanism for urgent revocation of dataset versions, and assistance with subject complaints. Without these, the agreement merely creates the possibility of recovering damages after a violation is discovered. The issue of transfer to third countries is also important if the broker effectively provides access from external infrastructure. The procurement decision should be recorded as a separate management act with any remaining uncertainties; it should not replace the legal assessment of initial and subsequent processing.

33. Public reporting: detailed enough, but not dangerous

Transparency of training sources is essential for individuals, regulators, copyright holders, and researchers, but the blanket publication of a list of all URLs risks the redistribution of sensitive information. Therefore, a differentiated approach to disclosure is needed: a public description of categories, periods, and rights mechanisms; a more detailed, secure set for the competent regulator; and selective evidence for an independent auditor, with legal guarantees of access [1][2][11].

"Summary of sources" and "list of all personal records" are not the same. The report should explain the method for categorizing sites, the scope and limitations of the list, the period of coverage, and version changes. Trust is enhanced not by decorative percentages, but by the reproducibility of what is included in the denominator, how the classification is confirmed, and how many objects lack full provenance.

It is proposed to publish separately the number of documented categories, the proportion of records with collection time, the proportion of sources with admissibility decisions, the number of processed applications for subject rights, and the sampling methodology. If the source data does not allow for estimating the denominator, the indicator is not published with fictitious precision; the lack of a statistical basis is explicitly stated.

In cases of actual harm, the type of consequences involved is specified: unlawful disclosure, outdated representation of the individual, hidden re-identification, or failure to enforce an objection. Data of applicants whose nature allows for their re-identification is not included in public examples without independent legal analysis.

34. Independent audit: subject of verification and limits of conclusion

An independent web scraping audit must examine four interrelated components: the legal justification for processing, the origin and authorization of sources, the actual operation of technical filters, and the effectiveness of data subject protection measures. None of these components can be replaced by the developer's statement of internal policy. However, the audit's conclusion does not have the force of a court decision or a binding order from a supervisory authority unless the law grants the auditor the relevant authority.

The audit target is a specific version of the collector, dataset, training run, and query processing interface. For each version, the auditor defines the period and scope of coverage, sampling methods, the number of sources reviewed, and access restriction classes. The statement "the company complies with the GDPR" without a specific subject of the audit does not correspond to proper methodology.

In situations of conflict of interest, when an auditor simultaneously advises the developer or represents a stakeholder, disclosure of financial and organizational ties, an independent risk assessment of the self-assessment, and, if necessary, a recusal are required. The lack of formal status of an accredited body cannot be disguised by official-sounding seals and certificate titles.

Conclusions are presented in a gradated manner: "confirmed by verified materials," "partially confirmed with limitations," "not established due to lack of evidence," and "specific discrepancies detected." This approach reduces the risk of unfounded accusations and protects interested parties' ability to submit objections.

An independent audit should document the purpose, scope, available sources of evidence, and the extent of verification before collecting data. Full access to a confidential corpus is not always necessary and may create a new risk of unlawful disclosure. Verification of selected manifests, filter reproducibility on a specially prepared test bed, log analysis, and model audit queries are possible. However, limited access requires an honest description of the scope of work: an external assessment of the model's outputs does not confirm the entire composition of the training set, and an analysis of the client's procedures does not prove that they were actually applied to each file. The report distinguishes between documentary evidence, a reproduced experiment, an indirect indicator, and information provided by a third party.

The structure of the final audit report includes the identified criteria, a list of supporting artifacts, a materiality classification, possible alternative explanations, the scope of the review, and remedial actions. A separate section is devoted to the expert's independence: remuneration received, relationships with the developer, work on the controversial architecture, the client's right to influence publication, and access to conflicting data. An organization that simultaneously develops a model, consults with its provider, and publicly evaluates its legitimacy must implement clear institutional barriers. The requirement for independence here does not imply the absence of any funding, but does require disclosure of interests and a mechanism to prevent influence on the conclusions.

35. Statistical sampling: why a few successful tests do not prove safety

In a large corpus, comprehensive manual review may be impossible. For independent evaluation, probability samples with a defined population and a known procedure for including elements are used. Rare but severe events must be separately considered: documents containing child data or

non-public medical history may constitute a tiny fraction of the overall corpus and still pose a risk to processing [1][2].

An optimal design might combine random sampling of all records with stratified testing of high-risk domains and targeted negative tests. These three components serve different purposes. The results of targeted stress tests cannot be statistically generalized to the entire corpus; the results of random sampling cannot be interpreted as a complete absence of rare violations.

When calculating proportions, the sample size and uncertainty must be reported. For example, if no direct identifiers are found in a random sample of 200 documents, this does not prove a zero probability of their presence in the entire population: the approximate upper limit of the two-sided Wilson 95% interval is approximately 1.9%. The calculation depends on the chosen statistical model and does not replace a legal analysis of the severity of the consequences.

The report should separately describe repositories inaccessible for review, excluded countries, classification model uncertainty, and potential selection bias. Recommendations for remediation should be linked to specific observations. The "percentage of legality" should not be calculated as a mechanical sum of compliance categories, which are not assigned numerical weights by law.

Statistical assessment of filter quality begins with defining the overall population. If only records from a single language, short pages, and no images were selected, the results obtained cannot be generalized to the entire international corpus. It is important to distinguish between precision, recall, and prevalence, as a low proportion of sensitive data detected may reflect both actual rarity and insufficient detection sensitivity. Auditing requires independent control labeling, and disagreements between experts must be documented. If records are stratified by source type, language, and age, overall population indicators are calculated using appropriate weights; a simple average across intentionally equal test groups can distort the results.

A zero-error detection rate in a sample cannot guarantee absolute error-freeness. Even for a simple, independent random sample, there is an upper confidence limit for the sample frequency, which depends on the sample size and the assumed probability. In the case of cluster dependence, repeated publications, and purposive sampling, the standard binomial model may not be applicable without adjustments. The researcher should explain the design before reviewing the results, not retroactively adjust success criteria, and separately publish rare classes where the data are insufficient for reliable inference. Furthermore, a filter accuracy result does not automatically grant legal permission to collect risky categories: the quality of control and the validity of the processing remain separate issues.

36. Filter performance metrics and classification error handling

The technical filter can detect names, contacts, document numbers, medical terms, and characteristics of minors. Its quality cannot be described solely by the match rate. Two-class validation requires sensitivity (recall), positive prediction accuracy (precision), false negative rate, skip rate by source group, and a description of the ground truth markup. Parameters may change dramatically when changing languages and genres.

A false negative could preserve sensitive data in the corpus, while a false positive could unreasonably exclude acceptable materials. The appropriate balance of thresholds depends on the severity of the potential harm and the intended purpose, but the threshold alone does not establish

legality. Decisions on exclusions should be reviewed by a human, taking into account the risks and, if necessary, with minimal access to the content itself.

Three sets are useful for auditing: a frozen control corpus with documented permissions, an adversarial set with rare and unusual identifier formats, and a representative sample of regular pages. It is advisable to record the algorithm version and test date, otherwise the results cannot be compared between releases. Creating dangerous test materials containing real, unnecessary personal information without proper justification is prohibited.

Additional validation concerns the bypass of exclusion rules after document transformation, such as OCR, translation, or segmentation. Preserving the original exclusion in the metadata ensures that the restriction is not lost during technical processing. The "exclusion successfully applied" metric should check the downstream path to the training version, not just the first filter.

37. Sanctions compliance: risk of secondary dissemination of false identity

A hypothetical scenario: in an open article, a person with a common surname is mentioned alongside a sanctioned subject. The automated collection saves the paragraph, and upon reuse, the model begins to associate the law-abiding person with the restrictive measures. In this situation, the subjects of research are the source of the original error, the method of its reproduction, the legal status of the information, and the measures to restore accuracy. No real individuals are described in this example.

The sanctions list must be verified against an official document and a competent source, taking into account identifiers, dates of inclusion and exclusion, possible namesakes, and status changes. A generative search, a link graph, or a name match are not independent proof of sanction status. The use of data about a specific individual in a financial decision requires a separate assessment of applicable data protection rules and procedural safeguards.

An independent audit links a "suspicious inference" to the provenance of specific fragments, if such a trace is available, and verifies whether correction is possible. Lack of access to model weights limits inference: it is possible to establish demonstrated behavior for selected queries, but not necessarily prove the causal contribution of a single document.

A practical recommendation is to prohibit the use of the model as a primary source of legal sanctions status, to record the time of review of the official list, and to establish a procedure for prioritizing the escalation of complaints regarding reputational damage. Such measures address both the protection of individuals and the quality of decisions by financial institutions.

For sanctions screening, the data source is critical not only because of potential personal identifiers but also because of the legal heterogeneity of events. Commercial databases can combine official sanctions records, mentions in news articles, legal disputes, affiliations with organizations, and name matches. A generative model trained on such data can transform ambiguous features into a smooth, but legally false, assertion about a person's status. Therefore, the system must consider the date, decision-making authority, individual identifiers, the validity of the restriction, the cancellation of the record, and the level of source verification. Training on aggregators that don't maintain a history of edits and confuse people with the same last name is especially dangerous.

A hypothetical case: a model reports that an entrepreneur is included on the EU sanctions list based on an old news article about another individual. An independent verification should not be limited

to demonstrating an incorrect answer. It is necessary to separately establish whether the corresponding entry exists on the official list on a specific date, whether the original source of the error can be identified, whether the error is recurring in different versions of the model, and whether a correction mechanism is provided. If an internal training set is unavailable, a causal inference regarding the error's origin remains hypothetical. The validity of the bank's subsequent decision is assessed according to industry requirements, separate from data protection. This distinction prevents both unfounded accusations by the developer and the dangerous practice of making sanctions decisions solely based on AI-generated data.

38. International search, extradition and criminal proceedings data

A hypothetical scenario: a national wanted notice, removed from public access, remains in the training building, and the model reproduces it without indicating the date or the subsequent refusal of the foreign court's extradition request. The danger lies in the confusion of different procedures: the national record, a possible international notice, an extradition decision, and the status in INTERPOL's information systems cannot be equated with one another.

When preparing a legal opinion, it is necessary to establish the primary procedural document, date, subject, jurisdiction, current status, and grounds for using personal data. Article 10 of the GDPR, if applicable, requires a separate verification of the legal regime. The authenticity of an international search is not confirmed by a generated model response, search results, or old newspaper publication [1].

For an independent assessment of a training corpus, it is important to determine whether refuted allegations develop into a long-lasting pattern. If it is technically impossible to eliminate all residual effects, the developer must assess the limitations of using the model in law enforcement or immigration scenarios, particularly where the conclusion could lead to imprisonment or denial of rights.

An expedited redress protocol is recommended: preserve proof of issuance, record the primary source and any judicial decisions, qualify sensitivity, limit risky downstream issuance, and provide individuals with a clear appeal route. Observatoire ARGA can develop a methodology or conduct an independent review within its available powers, but does not replace the formal review mechanism of competent authorities' decisions.

Information on international wanted lists is particularly sensitive to changes in legal status. The deletion of information by an international organization, a national court decision, a refusal to extradite, the termination of criminal proceedings, and a journalistic report on charges are not interchangeable documents. In the public training dataset, they may appear in different chronologies, and the model tends to summarize the most frequently repeated version. Therefore, for independent verification, a documentary timeline is first constructed, separating established decisions from assumptions, and then the quality of a specific AI response is assessed on a limited set of tests. Access to closed international police cooperation databases cannot be assumed based on the results of a public search.

An ethical and legal dilemma arises when publishing the audit itself. To demonstrate that the model has inappropriately linked an individual to a crime, the auditor may re-extend the original accusation. Limiting identifying details, confidentially releasing the full version to a competent recipient, and publicly releasing a statistical or anonymized conclusion are appropriate if the goals

can be achieved without causing further harm. In particular, the mechanical inclusion of full screenshots of search results and passport data in attachments should be avoided. This is not a blanket prohibition on the use of evidence, but a requirement for a proportionate, secure procedure, taking into account Articles 5, 6, 10, and 32 of the GDPR and applicable national law.

39. Suppliers and developers: responsibility for decisions with incomplete information

A developer purchasing a kit isn't always capable of independently conducting a comprehensive audit of millions of source pages. This objective limitation doesn't make the vendor's sole statement, "We comply with everything," sufficient. Proportionate due diligence requires risk-based questions, contractual guarantees, selective verification, retraction policies, and the ability to suspend further use.

It's important to distinguish between three types of errors: a known violation, an unresolved high probability of a violation, and a lack of sufficient evidence to support a conclusion. In a management report, these errors should lead to different management decisions. Calling any document deficiency an "established violation" is inappropriate; declaring it "compliant" due to the lack of a proven fine is equally unfounded.

The agreement must specify legally and technically enforceable mechanisms: notification of a claim, timely identification of affected versions, termination procedures, cooperation in evidentiary verification, distribution of information about subjects, protection of confidential materials, and the possibility of independent access within the limits of the law.

From a corporate governance perspective, the decision to accept residual risk should be made by an authorized body after receiving legal and technical information. This decision should not be delegated to a junior engineer simply because they run a scraper, or to a DPO as the sole owner of the business risk. Roles, review deadlines, and blocking conditions are defined in the management matrix.

40. Proposals of the Observatoire ARGA for EDPB practice and international dialogue

The first proposal is to establish a verifiable minimum provenance standard that does not require the publication of an unlimited list of personal URLs. Such a standard should cover source versions, collection purposes, categorical filtering, legal history, timelines, associated model versions, and query execution. The method of disclosure depends on the recipient: the subject, the supervisory authority, an independent auditor, or public reporting.

The second proposal emphasizes the two-pronged nature of opt-out effectiveness: preventing new learning and assessing the impact of completed learning. Formal exclusion of a domain from a crawler should not be automatically described as erasing information from the model's weights. Since machine unlearning capabilities vary, the legal standard should require honest disclosure of available means and evidence of their impact, rather than a technically unfounded promise.

The third proposal is to develop special safeguards for data on children, health, criminal offenses, and border/migration histories. It recommends risk-first definition of restricted source categories, quarantine of disputed fragments, limiting downstream use in high-consequence decisions, and

verifiable escalation of complaints. This proposal is institutional in nature; it does not expand the remit of the EDPB or an external organization.

The fourth proposal is to publish reproducible audit indicators: the size of the surveyed population, the sample quality, the proportion of unknown sources, filter sensitivity tests, the actual timeframes for resolving complaints, and a list of the independent researcher's limitations. A sound methodology is preferable to a "compliance rating" with fictitious weights.

The fifth proposal is to strengthen procedural safeguards: a reasoned response to requests, a mechanism for submitting additional evidence, protection of applicants' private information, separation of auditing from commercial consultations, and a clear indication of the limits of the opinion. The text of Guidelines 03/2026 under discussion serves as a starting point, not the existing new legislation. Proposals should be considered within the framework of the regulators' competence and applicable law [2][13].

Proposals to the EDPB should be formulated as specific clarifications to the draft Guidelines 03/2026, rather than a request to "strengthen protection" in general. The first area is verifiable criteria for sufficient transparency under Article 14 when collecting multiple external sources, including the notification language, source category descriptions, and the communication channel. The second area is guidelines for confirming mitigation measures for untargeted scraping, where the composition of visited websites changes rapidly. The third area is clear indications of what evidence of the practical effectiveness of filters can be taken into account in LIAs if significant false negatives are known in rare languages and special categories. However, the proposals should not contradict the legislative text or transform voluntary methodological forms into mandatory requirements without an adequate legal basis.

The fourth area is the compatibility of subject rights with the versioning of AI models. Guidance documents could encourage the transparent separation of deletion operations from raw data, training corpora, indexes, checkpoints, and model inference. The fifth area is supply chain auditability requirements, without disclosing personal data to auditors beyond what is necessary. The sixth area is the regular review of excluded source classes, taking into account vulnerable groups, legitimate objections from sites, and evolving technical methods. These proposals are aimed at reducing practical uncertainty; their admissibility and possible inclusion in the final EDPB text have not yet been determined. The report records them as the author's methodological position of Observatoire ARGA.

41. Conclusions, limitations, and future research agenda

The legitimacy of web scraping for generative AI is not determined solely by the availability of a page. It requires a separate assessment of the objectives, the basis under the GDPR, the categories of data, the expectations of individuals, alternatives, the protection of rights, and the actual flow of information throughout the entire technical chain. The publicity and innovative nature of the project may be relevant in the analysis of interests, but do not replace the mandatory legal prerequisites [1][2][3].

As of 22 October 2026, EDPB Guidelines 03/2026 are a draft submitted for public consultation until 30 October 2026. Therefore, the report does not ascribe to the draft the force of an effective stand-alone regulation and does not assume that its final version will be unchangeable. The current GDPR and the relevant provisions of the AI Act and copyright are assessed separately [2][11][18].

The proposed independent verification methodology requires evidence-based consistency from a specific source to model modifications and the exercise of subject rights. Due to the lack of access to the internal corpora of most developers, this approach does not guarantee full recovery of the training set in an external study. It is the documented boundaries of conclusions that distinguishes a high-quality expert assessment from an unprovable public accusation.

Further research includes an empirical assessment of the quality of filters in multiple languages, a comparative analysis of the application of Article 14(5)(b) in EU member states, methodologies for confirming the impossibility of reverse identification, and case law on the consequences of using improperly collected data in model development. Each area requires independent, verified sources, rather than a mechanical expansion of this document.

Appendix 1. Terminology map for independent opinion

The working opinion recommends providing definitions before assessing legality: crawler — programmatic page discovery; scraping — automated extraction and storage; training set — a finite collection of training records; fine-tuning set — subsequent tuning data; provenance — a verifiable history of provenance; data broker — an independent provider of a set; rights reservation — a copyright holder's indication of use restrictions for text and data mining under applicable conditions. These terms describe different legal and technical relationships.

The concept card includes the English term, the Russian form, the stages involved, an example of evidence, and the potential risk of erroneous confusion. Thus, "public availability" refers to the conditions under which information is visible, while "legal basis" refers to the legal basis for processing personal data. Publication on a website does not in itself confirm the existence of a basis under Article 6. "Anonymization" and "pseudonymization" are also not interchangeable: the latter usually preserves the possibility of reidentification with additional information.

When reviewing external company reports, a table of terms used and information on definitions taken from law, regulatory guidance, or internal standards should be included. Terms coined by the authors of the expert methodology are marked as proprietary to avoid assigning them official legal force.

Term	What needs to be confirmed	Typical mistake
Public source	Actual access and restrictions	Considered GDPR-free
Legitimate interest	Three conditions of Article 6(1)(f)	Equate to commercial gain
Anonymous dialing	Risks of allocation, connection and withdrawal	Equate to name hash
Removal	Specific data layers and verification	Promising to erase influence without testing
Independent audit	Scope, Method, and Limitations	Pass off as a regulator's decision

Appendix 2. Matrix of legal grounds and additional restrictions

This form applies to each individual processing step, not to the "AI project" as a whole. The first column specifies the operation, such as retrieval, storage, licensing of a dataset, training, or publication of the output. The second column specifies the controller or joint controllers determining the purposes and means. The third column specifies Article 6 of the GDPR and the

rationale for its applicability. The fourth column specifies specific limitations of Articles 9/10, child protection, transfer to a third country, and copyright.

If the source data consists of different classes, the map is compiled by risk group. A positive assessment for technical documentation without personal information does not apply to health reports or individual allegations. If the source category is unknown, the record is reassigned to "additional verification" status; lack of evidence is not considered proof of compliance.

Operation	Basis and additional rule	Solution/proof
Collecting public contacts	Checking Article 6; Balance of Expectations	LIA + category bans
Health information	Article 6 plus deletion of Article 9(2)	Legal review; if absent, block
Criminal record/offense	Article 10 and applicable law	Legal sanction or block
Downloading works	Applicability of DSM/TDM and redundancy	Rights reservation register
Transfer outside the EEA	Chapter V of the GDPR additionally	Transfer assessment

Appendix 3. Form LIA: Specification of Legitimate Interest

The form is opened before collection and assigned a unique number and version. The "interest" field must contain a specific legitimate purpose, a measurable desired outcome, the intended users, and the expected impact on individuals. An unacceptable example would be: "collect the entire internet for the sake of AI progress." A working example would be: "create a multilingual dictionary of legal terminology for a public information service without building individual profiles."

The categories of source websites, the collection period, the geographic scope, the expected type of personal data, the need for each category, and the potential recipients are then specified. The "missing facts" field is mandatory: if the source structure is unknown, the decision cannot be considered definitive. The next block establishes the connection between this interest and the stated purpose under Article 5(1)(b) of the GDPR and distinguishes the initial publication from the future use [1][5].

Required field	Responsible	Examination
Specific legitimate interest	Project owner	A verifiable goal, not a slogan
Categories of persons and materials	Data lead	Actual coverage
Scale and timing	Engineering	Real crawler limits
Alternatives	Product + privacy	Documented and compared
Admission decision	Authorized person	Date, version, reasons

When filling out the LIA, it's important not to leave descriptive fields in the declaration form. For each class of source, the operator must specify the specific benefit of the proposed processing, alternative methods, arguments against them, and possible negative consequences for people. For example, the desired goal is to improve the recognition of rare legal language. The entry may note the inadequacy of known licensed terminology corpora, but the conclusion must be supported by

comparative tests or a convincing plan for obtaining them. Citing a lack of available data does not prove that mass collection of public discussions with biographical information of participants is permitted.

An independent LIA review should include negative decisions. If at least one category of source was not rejected, it is useful to check whether the opposing arguments were even considered. The proposed form includes a "counterevidence" field: what fact could have changed the decision. For example, the discovery of a high proportion of archival data on criminal prosecutions would require a review of the pre-launch clearances. The LIA date, responsible person, and justification version are linked to the actual crawler configuration so that the review can determine whether restrictions were in effect at the time of data extraction. This allows for the traceability of the organizational decision and does not confuse internal approval with the approval of the regulatory body.

Appendix 4. LIA Form: Necessity and Less Intrusive Alternatives

A comparison is conducted before selecting a mass scraping method. The following are included: a licensed dataset without personal data; an aggregated statistical corpus; synthetic samples; limited collection by domain; exclusion of personal profiles; metadata minimization; and archive depth reduction. For each option, the achievability of the specific task, the expected quality of the result, the impact on rights, and the reliability of the initial estimates are assessed.

The justification cannot simply be the phrase "alternatives are technically impossible." For example, a sample of a limited class of industry-specific texts may satisfy the needs of a specialized assistant and reduce the risk of extracting biographies of others. If the developer rejects it due to a decrease in quality, a control experiment and the magnitude of the change, as well as the sample size and characteristics, must be specified.

Alternative	Examination	Conclusion
Licensed anonymized texts	Comparison of quality on hold-out sets	Acceptable/inadequate and why
Collection of only thematic domains	Loss of target language coverage	Controlled restriction
Synthetic examples	Leakage, quality, representativeness	Application with protective conditions
Pre-exclusion of personal profiles	The percentage of omissions in the test	Approve/Rework Filter

Appendix 5. Balance of Interests: Model of Harm Caused and Prevented

The matrix does not introduce artificial legal scoring. It is intended to provide a structured argument. It describes the risks of each type: loss of control over publication, unwanted disclosure of information from the model, re-identification, stigmatization, mischaracterization as a criminal or sanctionable offense, lack of transparency, and obstacles to the exercise of the right to object. For each, typical affected individuals, likely mechanisms of harm, mitigating measures, and residual uncertainty are provided.

A separate clause concerns the subject's reasonable expectations. An individual professional profile and a personal message on a forum may be equally indexed, but the expectations for future use differ. The justification does not provide an abstract legal opinion, but rather the facts of the context, scale, and future goals. The priority of protecting the child's interests must be given special consideration [1][2].

The conclusion specifies which measures have been confirmed by the test, which are still planned, and who makes the final decision. If a significant nonconformity cannot be resolved, the form stipulates that work with the affected category will be blocked. A lack of documentary evidence is recorded as an unresolved issue, rather than disguised as "low risk."

Appendix 6. Admission and prohibition of Internet sources

The domain list should not be reduced to a technical allowlist. Each source is assigned attributes: site category, access authorization, publicity, special types of personal data, presence of minors' data, robots.txt/CAPTCHA and other explicit restrictions, known contractual terms, publication period, and possible content rights. A source may be allowed for technical metadata monitoring, but prohibited for full-text preservation.

Four statuses are offered: approved with conditions; temporarily restricted; blocked; and pending review. Each status is assigned a decision owner and a re-evaluation deadline. Changes to the site's architecture, accessibility, or conditions do not automatically inherit the previous approval status. When performing mass collection, it is necessary to verify that the crawler is actually using the latest approved version of the source list.

Sign	Example of a management decision
A forum with potential personal medical histories	Preliminary blocking; legal review
Official open set of technical terminology	Admission after checking the license and data
Resource with a technical objection to collection	Exclusion from automatic bypass
Domain with unknown status and child data	Quarantine until a reasoned decision is made
Old court archive with data on crimes	Separate verification of Article 10 and access rights

Appendix 7. Source Intake Sheet: Primary Source Card

The card records the decision date, the original URL or restricted identifier, the resource owner, access conditions, language, jurisdiction, subject matter, license source information, personal data class, and the justification for the need. When processing particularly sensitive materials, the direct URL may be stored in a closed registry, while the public version may use a secure source pseudonym.

Example of a conditional entry: SOURCE-LEGAL-2026-001, "official public documents with legal terminology," access without personal registration, robots.txt allows only some sections, pages with personal complaints and search results for individuals are excluded. Decision: "approved for a limited corpus of technical texts," review period in three months, owner - data governance manager. The entry does not imitate information from a real website.

Card field	Required entry
Source ID and version	Stable identifier
Date and route of access	Factual verification, not speculation
Legal status	GDPR, content, special restrictions

Card field	Required entry
Excluded sections	Specific patterns and paths
Level of evidence	Initial inspection/test/external audit

In practice, a primary source card begins with a precise record of the domain and type of material, but this is not sufficient to establish the origin of a specific record. A distinction must be made between the host site, the original author, the aggregator, and the subsequent republisher. For example, a journalistic publication might quote a court ruling, while another portal might reproduce it with abbreviations. If the model is trained on the latter, the error message should include the path to the original document, but the lack of such a path should not be compensated for by guessing based on textual similarity. The card records the retrieval URL, date, category, known original sources, level of verification, and restrictions on copy distribution.

To control large-scale collections, a two-step verification process is useful: automated initial categorization of domains and selective human review of categories where errors can be costly. It's important to record not only approved sites but also the reasons for rejection. If a site was excluded due to explicit objection to scraping, the system requires a mechanism to update the ban when the URL or mirrors change. However, a technical robots.txt signal is no substitute for a full legal review; its absence does not mean all types of processing are permitted. If there are any discrepancies, the card is moved to "suspended" status until the circumstances are clarified.

Appendix 8. Dataset Manifest: Minimal Structure of Proof of Origin

The unified manifest includes the Dataset ID, version, created_at, purpose ID, legal basis reference, source category, crawl job ID, source window, content type, PII-filter version, retention rule, transfer policy, downstream model runs, and subject declaration status. Version changes should be append-only in the log, not overwritten without a trace. There must be a way to prove which file was used in a specific training run.

Пример записи, предназначенный только для иллюстрации: DATASET-DEMO-07; версия 2.1; технические публикации за определённый период; общий корпус 18 000 условных документов; исключены личные профили, спорные сведения в карантине; контрольный журнал содержит 12 отклонённых URL; дальнейшее обучение допускается только после обновления фильтра. Все цифры здесь учебные, не описывают операционную статистику какой-либо компании.

Контроль полноты manifest не ограничивается заполняемостью. Аудитор сравнивает выборку строк с логами extraction, каталогами raw/clean, хэшами и записями использования в model registry. В случаях отсутствия связи вместо восстановления «по памяти» присваивается статус traceability gap с оценкой влияния на выводы.

Идентификатор	Назначение	Обязательное доказательство
Crawl Job ID	Процесс извлечения	Конфигурация, логи, дата
Dataset Version	Состояние набора	Хэш списка объектов
Model Run ID	Использование в обучении	Манифест загрузки и обучения
Rights Ticket ID	Исполнение запроса	Решение, действие, результат

Идентификатор	Назначение	Обязательное доказательство
Exception ID	Ограничение в источнике	Причина, владелец, срок

Приложение 9. Проверка поставщика уже собранного корпуса

Покупателю рекомендуется запросить цепочку происхождения по категориям и представительную выборку данных, не получая лишние персональные сведения. Запрашиваются география и периоды сбора, допустимые цели повторного использования, legal basis, статья 9/10, правовой статус контента, история возражений, удаление данных по жалобам, список субподрядчиков и содержание security controls. Отдельно проверяется, способен ли поставщик определить объём ранее полученных и уже удалённых данных.

Внутренняя шкала проверки не должна позволять компенсировать отсутствие правового основания большим количеством технических сертификатов. Юридически блокирующие недостатки отмечаются отдельно от обычных эксплуатационных несоответствий. Если поставщик не может предоставить сведения из-за коммерческой тайны, стороны могут организовать контролируемый просмотр доказательств независимым экспертом; это не должно вести к раскрытию данных без правового основания.

Вопрос поставщику	Доказательство	Статус ответа
Кто являлся контролёром при сборе?	Карта ролей и договоры	Подтверждён/оспаривается/неизвестно
Как исключаются запрещённые сайты?	Правила crawler и журнал тестов	Подтверждён/неподтверждён
Как исполняются возражения после продажи?	Процедура уведомления покупателей	Реализовано/не реализовано
Есть ли данные статьи 9 или 10?	Протокол классификации	Наличие/отсутствие/неизвестно
Можно ли проверить конкретную версию?	Manifest, хэши, dataset lineage	Полный/частичный след

The dataset provider should not be asked the general question "is the dataset GDPR compliant?", but rather a series of verifiable queries. The buyer requests website categories, time periods, the origin of the licenses claimed, the grounds for further processing, and a description of the procedures that exclude special categories, child data, and criminal prosecution information. For some questions, an aggregated answer may be sufficient if detailed disclosure would itself threaten the rights of third parties. However, any answer must be dated and relevant to a specific version of the dataset; an old audit certificate for a different collection does not confirm the properties of the new delivery.

If a vendor claims "all records are anonymized," it's reasonable to request a re-identification threat model and a limited linkage test. Sometimes identification is possible not just from a single field, but from a rare combination of content and external data. If independent testing is refused, the risk should be assessed separately, rather than turning a commercial privacy dispute into a presumption of violation. A procurement decision considers three outcomes: sufficient evidence, remediable gaps, and significant uncertainty. In the latter case, it is methodologically justified to withhold approval for certain categories until the obstacles are resolved.

Appendix 10. RACI and delineation of powers of participants

A single team should not simultaneously determine the economic objective, recognize the legal basis as proven, verify its own filters, and approve an external report. For a web scraping project, the essential roles are those of the processing owner, data lead, DPO/privacy counselor, security, product, subject rights specialist, and independent auditor. The DPO advises and oversees, but does not automatically become the owner of the commercial solution.

The RACI model: the project manager sets goals and is responsible for resources; privacy counsel verifies legal grounds and risks; engineering is responsible for limits and version integrity; security is responsible for access and incidents; the rights team is responsible for processing requests; and an internal committee decides whether to accept residual risk. An independent auditor evaluates a limited set of evidence without replacing organizational management.

Action	Responsible	Accountable	Inspector
Purpose and scope of collection	Product	Authorized Manager	Privacy counsel
LIA and DPIA	Privacy + data lead	Controller	DPO (Delivery Performance Objective)
Data filters	Engineering	Data governance owner	Independent tester
Rights of subjects	Rights team	Controller	Privacy counsel
Publication of the assessment	Observatory audit lead	Editorial organ	Independent review

Appendix 11. Public information notice under Article 14 of the GDPR: draft structure

An organization that obtains personal data from publicly available internet resources to train a generative model must prepare the information in plain language. This sample structure is not a universal, official template: the specific mandatory information is determined by Article 14 of the GDPR, the actual source, and the characteristics of the processing. The notification includes the name and contact information of the controller, the purposes and grounds for processing, categories of information, types of sources, recipients, timeframes, possible transfers, rights, and the procedure for handling [1][2].

Example of content for adaptation: "We use limited categories of publicly available materials to develop a text processing system. In some cases, these materials may contain information related to identifiable individuals. We apply rules for screening, source exclusion, and data cleansing. You can contact us through the specified channel to request information about the processing, correction, deletion, restriction, or to object in cases provided by law. We consider each request taking into account the actual processing and applicable law."

Actual domain categories and collection periods must be specified, rather than the abstract phrase "all publicly available sources." When applying the exception to individual notification under Article 14(5)(b), the public notice must remain readily accessible and contain the information required by law. Before publication, it is necessary to verify that the list of sources does not reveal new private information. The presence of such text does not relieve other obligations.

Notification block	What does the reviewer check?
Controller and contacts	A real opportunity to appeal
Purpose and basis	A specific, not a general phrase
Sources	Categories, period, restrictions
Rights	Working procedures and response times
Recipients and transfers	Correspondence to real architecture

Appendix 12. Justification for Article 14(5)(b): Documented Procedure

The controller describes the number and categories of potential subjects, the time scale of the material, the availability of contact information, the feasibility of direct notification, the anticipated costs, and the additional risks of identifying individuals for the purpose of sending the message. It then compares the lack of individual notification with the potential rights implications, and identifies alternative methods of notification and additional security measures [1][2][14].

An example of a conditional assessment: a library of historical forum discussions spanning 15 years may contain references to several hundred thousand individuals without current contact information. Creating a list of addresses may require additional bulk extraction, increasing the risk. This circumstance does not exempt the user from general protections: a published policy, restrictions on personal information, an accessible reporting channel, and periodic review of the rationale are required. The final conclusion depends on the specific dataset.

Форма содержит подпись ответственного за цели и документированную правовую оценку, но юридическое основание определяется нормой, а не количеством подписей. Пересмотр проводится при возникновении новых средств связи, выявлении чувствительных материалов или изменении масштаба сборки. Решение с неизвестными параметрами маркируется как предварительное.

Форма для статьи 14(5)(b) должна отражать не только число потенциальных субъектов, но и затраты на индивидуальное извещение, доступность контактной информации, реальные препятствия установления лиц и тяжесть последствий для их прав. Простая оценка «данных слишком много» не раскрывает критерий несоразмерности. Полезен сценарный расчёт по отдельным категориям: например, для профессиональных регистраций адресов электронной почты может быть достаточно, тогда как для архивных постов без надёжных контактов ситуация отличается. Нельзя автоматически переносить вывод по одной категории на весь универсальный корпус. При этом сам сбор адресов исключительно ради уведомления способен создавать новый самостоятельный риск, который также оценивается.

Если исключение применимо, проект требует мер, защищающих законные интересы субъектов. Среди них может быть доступное публичное уведомление, описание источников, механизм обращений, эффективные средства предотвращения новой обработки после обоснованного возражения и отдельные правила чувствительных категорий. В акте должны быть представлены не только намерения, но и результат проверки: открывается ли уведомление без регистрации, понятен ли язык, поступают ли заявки, исполняются ли они в установленный срок. Исключение из индивидуального информирования не означает освобождения от остальных принципов GDPR и не аннулирует права субъекта.

Приложение 13. DPIA: карта угроз правам и свободам

Первый лист описывает операции, цели и необходимость. Второй — категории лиц, объём данных, правовой режим и сроки. Третий — угрозы и имеющиеся меры. Четвёртый — остаточный риск, обоснование решения и необходимость предварительной консультации по статье 36 GDPR в соответствующих обстоятельствах. Оценка должна осуществляться по фактическим признакам, а не по универсальной формуле «масштаб большой — нарушение доказано» [1][16].

В учебном примере генеративная система собирает научные публикации, но дополнительно сохраняет неанонимизированные обращения пациентов на медицинском форуме. Потенциальный вред включает раскрытие здоровья, ошибочные медицинские выводы и невозможность контроля дальнейшего распространения. Предлагаемые меры: запрет домена и аналогичных источников, уничтожение уже собранных копий, аудит восстановления сведений, перепроверка связанных training runs и информирование о доступных правах в установленном порядке.

Риск	Механизм	Предлагаемая защита
Ненужные сведения о здоровье	Автоматический обход форума	Доменные исключения; фильтры; карантин
Недостоверное обвинение	Старый новостной корпус	Временные атрибуты; проверка статуса
Повторная идентификация	Объединение таблиц	Limit linking; threat model
Impossibility of objection	No source connection	Rights ticket + dataset lineage
International access	Remote contractors	Chapter V and Access Control

Appendix 14. Stop rule for data under Article 9 of the GDPR

The "stop-isolate-verify-document" rule is applied when special categories are likely to emerge. First, the category and the minimum required evidence of detection are recorded. The affected batch is then transferred to a closed repository without access to training. A lawyer verifies whether a specific exception under Article 9(2) exists, provided there is a legitimate basis under Article 6; if no appropriate exception exists, the data is excluded and deleted according to the approved procedure.

It is not recommended to share sensitive documents with a wide range of employees for "clarification," as the test itself may increase the risk. The internal report indicates the percentage of positives and the test's limitations. If the filter only detects a portion of such information, the inability to assess its completeness should be reflected in the conclusions and follow-up plans.

Stage	Who is acting?	How to fix
Discovery of probable data of Article 9	Data governance	Risk event ID, attribute type
Quarantine	Security / engineering	Access restrictions and logging
Legal qualification	Privacy counsel	Art. 6 and specific Art. 9(2)
Remediation	Data engineering	Remote versions, retraining risk
Re-check	Independent reviewer	Sample and conclusion limitations

Appendix 15. Stop Rule for Information on Criminal Offenses

Data processing under Article 10 of the GDPR requires a separate legal analysis and should not automatically be included in the general procedure for ordinary public documents. For a suspicious record, only the necessary characteristics are retained: official or unofficial source, type of suspected legal event, date, status, possible identification conflict, and jurisdiction. Whether an organization is authorized to process this class of data under applicable law is determined separately [1].

The conditional entry "person linked to investigation" should not be converted to the label "conviction" without a valid source document. Police diffusions, wanted notices, extradition requests, and court verdicts have different legal status and evidentiary value. If there is insufficient evidence to support the use of the data in training, it is reasonable to suspend the inclusion of such a fragment.

The remediation map contains a comparison of the erroneous and correct versions, evidence of dissemination, a decision to prohibit use, and methods for limiting model inferences about a specific individual. External communications should not publish additional information about the applicant without independent justification.

Appendix 16. Child Protection: Domain and Content Exception Rules

The team identifies sources where children's information is expected, including school pages, youth discussion groups, family photo galleries, and assistance to minors. The default action in high-risk categories is to refuse collection until further review, rather than automatically enabling it and then anonymizing it. When working with age-based indicators, unnecessary additional collection of personal documents should be avoided [1][2].

The verification protocol includes a set of safe synthetic examples, an assessment of filter bypasses in different languages, an OCR evasion test, and short multimodal captions. The report displays errors limited to the test set. It should not be assumed that a large language model can reliably determine a person's age from the text of every online post.

Verification indicator	Question
Preliminary exclusion of sources	Are there any classes of sites that are not visited by crawlers?
Erroneous omissions	What childhood references were not noticed?
Subsequent use	Are the following training runs excluded?
Rights of parents/representatives	Is there an active channel of appeal?
Residual uncertainty	Are the classifier's boundaries documented?

Appendix 17. Procedure for processing a data subject's request

Each request is assigned a number, a category of requirements, a date of receipt, a contact for a secure response, identity verification only to the extent necessary, a linked dataset ID, and a responsible team. A search is then performed using valid identifiers, the legal obligation is analyzed, measures are determined, and the response is documented. The card automatically records

the one-month reference date under Article 12(3) of the GDPR, a note indicating timely notification of the requester in the event of a reasoned extension, and the date of the final response. An extension is possible for a maximum of two additional months, subject to the established conditions [1].

The requester's interface should not require precise knowledge of the internal technical structure. A query containing the name and link to the problematic publication may serve as a starting point for a search, but the risk of a false match requires confirmation before sharing the information with an outside party. If the controller cannot securely identify the individual, they explain the obstacle and the legal means of refining the query.

Step	Documented action
Reception	Number, request type, date
Identification	Minimum sufficient verification
Search	Sources and related versions
Legal assessment	Applicable Article and Exceptions
Remediation	Raw/clean/index/model, degree of confirmation
Answer	Reason, date, method of challenging

The request workflow is built around a log of restricted access commitments. Upon receipt, the date of receipt, the subject of the request, the type of proposed right, the required identity verification, and the list of affected controlled systems are recorded. The legal applicability of the relevant right is then determined: requests for deletion, restriction, correction, and objection are not subject to the same conditions. A request should not be denied outright simply because it is technically impossible to prove the removal of an impact from all model weights; in the presence of controlled source or indexed records, permissible measures must be assessed separately.

When interacting with brokers, it's important to consider routing: who the request is forwarded to, who makes the decision, what confirmations are returned, and whether the buyer can block re-downloads from the old version. Requests from individuals with identical names should not result in the disclosure of another individual's data. The response protocol specifies only the actions verified, any remaining restrictions, contact information for subsequent disputes, and information on possible remedies under applicable law. This procedure is also useful for assessing the quality of internal compliance: it allows for measuring response times without publishing sensitive complaint details.

Appendix 18. Map of removal execution at different levels of the model

The technical action "delete data" is broken down into levels: original scrape, cleaned document, training dataset, temporary cache, retrieval index, evaluation set, logs, and model weights. For each stage, the presence, the possibility of technical deletion, confirmation of execution, and the remaining constraint are recorded. It is important not to claim successful "forgetting" simply because a single file has been deleted.

In the training scenario, a disputed record is found in the search index, but the origin of the model weights does not rule out earlier use. The team removes the record from the index set, prohibits its

re-retrieval, investigates its relationship with training versions, and separately evaluates methods for mitigating the risk of issue. The report reports what is confirmed and what cannot be determined.

Data layer	Potential action	Proof
Raw objects	Remove or isolate	Backup log and verification
Curated set	Exclude entry/version	Dataset manifest before and after
Retrieval index	Delete vector and fragment	No-result test
Model weights	Rate retraining/unlearning	Documented experiment
Model output	Filter + control test	Behavior on agreed requests

The deletion technical map links each request to different storages and models. For raw data, the deletion or restriction of a specific record is confirmed; for a prepared corpus, it is excluded from the version and future builds; for retrieval, it is reindexed; for fine-tuning, it is reviewing the checkpoint; for a parametric model, it is testing the retrieval and evaluating available measures. Furthermore, the storage of evidence of execution must itself have a basis, a duration, and appropriate security measures. It is not permissible to permanently retain a full copy of a disputed record in an open ticket for the sake of deletion confirmation.

The report must distinguish between the result and the prediction. The entry "planned to be excluded in the next training run" does not equate to a confirmed removal from the model currently available. Conversely, the failure of a specific data extraction test does not prove that the model will never be able to generate the data using another method. Restricted formulas are recommended: "removed from the specified index on the date," "excluded from the set of version X," "no replication was detected in test protocol Y," and "no effect on weights was determined." For each formula, the identifiers of the tested system and the test conditions are preserved. It is especially important not to claim the actions of independent third parties when the organization does not control their copies and has not received confirmation.

Appendix 19. Audit protocol of actual crawler operation

The auditor requests configuration snapshots, parser versions, link crawling rules, the user agent used, schedule, speed limits, and a list of allowed and blocked domains. A random sample of visits is taken from the network logs to determine whether actual operations match the approved scope and how exclusion rules are processed. If the logs are incomplete, this is indicated as a limitation of the audit.

The test is conducted in a controlled environment or on legally accessible websites. Bypassing CAPTCHA, authentication, or technical protection on another website without a valid reason for the sake of auditing is prohibited. The accuracy of the procedures can be verified on your own simulated websites containing artificially generated data and various restrictions.

Test	Expected result
robots.txt disallow for control path	Crawler does not fetch the page
Page outside the allowlist	URL rejected
Page with prohibited content class	Batch quarantined
Rule changes in progress	The version on the task date is applied

Test	Expected result
Network/logging failure	The error is recorded, not hidden.

Appendix 20. Data protection model in the contractor chain

An example of an architectural map: the project owner approves the goal and set of sources; the scraper contractor executes instructions and maintains extraction logs; the cloud operator stores a limited collection; the labeling company receives a selected subset; the developer trains the model; the rights of the subjects are processed by a contact assigned by the controller. For each party, the actual roles, access, region, operations, and retention period are established [1][4].

The auditor considers not only the formal existence of a DPA but also the fulfillment of contractual terms: whether there is a subcontractor log, a policy for deleting copies after contract termination, the ability to return/delete objects, and the ability to assist with a subject's request. An opaque transfer path is especially dangerous for data of special categories or criminal records.

It is recommended to conduct a "revocation of external access" test: terminate the former contractor's access, confirm the deletion of their working copies, and verify that training does not rely on unverified third-party archives. A positive result from one test does not confirm the absence of other copies, so the test limitation is clearly stated.

Appendix 21. Copyright and Text and Data Mining Rights: Separate Check

The legal content card must distinguish between the existence of a protected work, the applicable exception or license for the relevant operation, the lawfulness of access, and the presence of a properly expressed reservation of rights. According to Article 4 of Directive 2019/790, the exception for TDM in the context of lawful access may be based on the lack of an appropriate reservation of rights for the copyright holder. The specific implementation and the effect of other provisions are examined separately [18].

During the audit, the resource type, the method of expressing restrictions, the validity period, the selection rule, and the link to the GPAI training content summary are recorded. A legal assessment of personal data is conducted in parallel and is not replaced by the TDM conclusion. It is advisable to separately state in the conclusion: "copyright use conditions confirmed/not confirmed" and "reason for personal processing confirmed/not confirmed."

Source	What to document	Possible conclusion
Licensed Corps	License scope and downstream rights	Use within the conditions
Open site	Legal access, TDM reservation	A separate assessment is required
Someone else's closed base	Agreement, technical limitations	Do not assemble without permission
Publication with personal information	GDPR plus content rights	Two parallel analyses

Appendix 22. The AI Act, GDPR, and Disclosure of Training Content

The matrix links the public summary under Article 53 of the AI Act with the provider's internal documents and privacy. A publicly accessible summary of content is not required to disclose

complete, identifiable documents, but must be sufficiently informative in accordance with the applicable template and legal requirements. Full technical documentation has a different audience than a publicly available publication [11][12].

In a working example, the company lists "publications from open scientific websites, specialized industry articles, and licensed materials" in its public summary. Within the organization, each category is assigned a record ID, time period, collection rules, the absence or presence of personal data, an exclusion policy, and a link to models. Without an internal link, a public formula does not constitute proof of actual provenance.

Public facility	Internal evidence
Training Content Category	Dataset ID and category source
Collection period	Crawl logs and timestamps
Copyright Policy	Rights reservation register
Legitimacy of personal processing	LIA, DPIA, and related records
Availability of subject rights	Tickets and documented responses

Appendix 23. Map of cross-border access to personal data

The controller maps physical storage locations, cloud regions, administrative access, subprocessors, and backup routes. For each potential international access, it determines whether a legally significant transfer exists and, if so, which mechanism under Chapter V of the GDPR is used and what measures are necessary. The international transfer regime complements, and does not replace, the general review of Article 6 [1].

The practical audit includes checking service accounts, external annotation handlers, diagnostic logs, and support teams with remote access. A separate question is whether personal data from user prompts could end up in a monitoring system not described as part of the initial training pipeline.

Element	Security Question
Cloud region	Where is the actual copy located?
Administration	Who has access and from which country?
Subprocessing	Is there a resolution and change log?
Legal mechanism for transfer	Is it installed and applicable?
Transfer of disputed data	Is temporary blocking and correction possible?

Appendix 24. Plan for statistical verification of source prohibitions

The population is defined as all elements of a specific corpus version to which the collection rules apply. A random sample with a fixed seed and a separate targeted high-risk sample are formed. For the former, an estimate of the proportion of specific errors with an uncertainty interval is acceptable; for the latter, inference is limited to the tested classes and cannot be extrapolated to the entire population without weights.

Case study: out of 200 randomly selected documents, two fall into categories requiring manual review. The observed proportion is 1%. Wilson's two-sided 95% confidence interval is approximately 0.27–3.57%, assuming the appropriate sampling model is applicable. These figures are not estimates for a real corpus and do not imply that 99% of documents are legitimate: the classifier only detects the selected feature, while legal legitimacy includes other requirements.

Variable	The importance of the case study
Random sample size	200
Controversial categories found	2
Observed share	1%
Wilson interval (95%)	≈0.27–3.57%
Forbidden output	“99% of the corpus is legal”

The statistical plan should include a probability of selecting each record or clearly explain that the sample is purposive. A quasi-random sample of "fifty beautiful examples" is suitable for illustrating the problem, but not for reliably inferring the proportion of violations in a multi-million-strong dataset. When underrepresented languages pose a high risk, stratification is advisable: large language groups are assessed separately from rare ones, and sensitive website types have their own control sets. If the results are subsequently aggregated, the weighting of each group should be taken into account by the study design. Uncertainty about the proportions of groups limits the ability to calculate the overall frequency.

The protocol records the population size, sampling unit, random selection method, error criterion, instructions to annotators, dispute resolution procedure, and uncertainty calculation plan. It is important to measure both the erroneous exclusion of safe data and the retention of prohibited content. For legal analysis, the latter error often has more serious consequences, but it is impossible to automatically declare every detected error as illegal processing: it requires knowledge of the factual basis, context, and measures taken. The auditor must distinguish between the statistical result, the technical explanation, and the legal conclusion.

Appendix 25. Personal Data Detection Test Matrix

The filter is evaluated across multiple languages and source types: plain text, HTML with custom markup, translated fragments, OCR, reposts, and structured forms. Training and test examples should be separated, and artificial or specially authorized test records are preferred for external audit. Metrics are calculated using known markup, taking into account class balance and the proportion of cases with uncertain classification.

The worksheet includes the filter version, data set, test volume, true positives, false positives, false negatives, precision, recall, and limitations. If 92 sensitive samples are detected out of 100 confirmed, the recall is 92%. However, this result alone does not indicate precision unless false positives are identified among other documents. Similarly, success in English does not indicate success in French or Russian.

Test class	The main risk	Control
Legal names and namesakes	Misattribution	Context and official status

Test class	The main risk	Control
Free medical text	False negatives	Semantic verification
Translated pages	Loss of sensitivity	Multilingual hold-out
OCR of documents	Hidden numbers and dates	Recognition error test
Controversial data of children	Age ambiguity	Blocking sources and review

Appendix 26. Independent audit program for 20 working procedures

The audit begins with agreeing on the scope: which model version, which corpus, and which countries are covered. Then, legal access to materials and confidentiality restrictions are agreed upon. Separate tests verify the following: processing purpose; LIA; DPIA; data category; list of domains; collection prohibitions; source and time; versioned manifests; contracts with brokers; supervisory roles; subject rights; storage and deletion; international access; crawler configuration; filter accuracy; downstream traceability; memorization risk; copyright disclaimers; GPAI public summary; and the independence of the report signatory.

For each procedure, the document request, the evidence obtained, the test performed, the result, and any uncertainty are recorded. Unavailability of the log should not result in a result being based on interviews. The report distinguishes between the verified absence of a clear violation in a given sample and evidence of compliance of all actions with the regulatory regime. In the event of a legal dispute, cautious formulas are used and the opportunity to present the position of the participant being audited is provided.

Procedures	Type of evidence
1-4: goal, LIA, DPIA, categories	Solutions, assessments, data maps
5–9: domains, bans, provenance, brokers	Logs, manifest, contracts
10–14: roles, rights, storage, translations, crawler	Access, tickets, test downloads
15–18: filters, models, leakage, copyright	Test reports, rules, and code
19–20: Publicity and Independence	Conflict reports, declarations

The independent audit program is divided into a review of design decisions and actual implementation. During the design phase, approved objectives, architecture, LIA, DPIA, controller map, vendor agreements, and domain exclusion policy are collected. During the operational phase, logs, actual crawler configurations, filter versions, transformation chains, evidence of complaint response, and periodic leak tests are analyzed. The third phase verifies the organization's ability to resolve the detected error, not just demonstrate the existence of a written procedure. It is recommended to define the type of evidence and sufficiency criteria for each item in advance to avoid changing the audit standard for the sake of a convenient outcome.

The review can be conducted in a secure environment with independent observers and minimal data copying. Lack of full access should be reflected in the report, including the types of data the expert was unable to verify. If the client refuses to disclose key logs under the pretext of confidentiality, this does not automatically prove a GDPR violation, but it does limit the scope of the positive

review. It is useful to break the report down into confirmed findings, insufficient data, significant deviations, and suggestions for improvement. In the event of a conflict of interest, an independent reviewer and a person authorized to decide on publication are appointed.

Appendix 27. Evidence Chain: example of movement of one document

Conditional document D-2026-001 was found on an authorized source on a specific date. The crawler log contains the time and URL, the raw object is hashed, the cleanup process excludes contact strings, and the manifest records the identifier of the resulting fragment. The training run references the catalog version. If a complaint is received, the rights ticket includes a check of the linked document and all subsequent versions.

The proof must allow us to answer the following questions: did the object exist at the time of loading; how did it change; on what basis was it extracted; was it included in training; what data from it was preserved; what corrections were made. Without such a trace, it is impossible to establish the source of an individual generation with the reliability sometimes promised by marketing "data lineage" models.

It's important to distinguish between proven digital connection and the causality of generative output. Even if the disputed text was part of the training set, the specific wording may arise from other data. Conversely, the absence of a discovered training record due to incomplete access does not prove its historical absence.

Artifact	Integrity control
Original page	Time, URL, access context
Raw object	Hash, limited storage
Cleaned fragment	Version ID, filtering
Training manifest	Connection with run ID
Rights ticket	Actions and final check

Appendix 28. Case Study: Wrong Sanctions Attribution

Simulated situation: A public announcement links a person with the same last name to a sanctioned individual. The generative system reproduces the dubious connection without specifying the source, date, or level of certainty. This example does not refer to a specific person or case. The independent verification begins by recording the model output and the precise conditions under which the result was obtained, then verifies the public source and any possible official documents.

Next, the probability of the presence of an erroneous fragment in the assembly is determined if the source data is available; the model's use in the KYC/AML process is assessed separately. It is important to distinguish the technical conclusion "for this request, the model has identified a link" from the legal conclusion "the individual actually has a sanctioned status." The choice of action depends on the origin of the error, the degree of public dissemination, and the subsequent consequences.

Corrective action plan: temporary restriction of risky downstream release, verification of the official list, adjustment of the recruitment process, notification of relevant departments, a reasoned response

to the individual, and subsequent testing monitoring. No step should create another publicly accessible copy of the disputed personal information without a legal basis.

For a training case of sanction misattribution, we need to create a set of artificial records that don't replicate real private accusations. Person A and Person B have similar last names but different dates of birth and jurisdictions. A news article mentions an investigation against Person B, while the sanction in the official source refers to a different entity. Through several aggregators, the text is converted into an abbreviated record with an indefinite pronoun. The model responds that Person A is under sanctions, although the original list doesn't confirm this. The protocol separately evaluates the quality of the response, possible mixing of records, and the lack of evidence that the error arose specifically from the training document.

Remediation in such a case involves checking official, up-to-date registries, clarifying identifiers, modifying the downstream decision-making process, and establishing a channel for correcting the record in the controlled corpus. If the banking system uses a model for verification, an independent assessment of the requirements for the financial institution and human verification is necessary. This case study demonstrates the difference between acceptable research into the mechanism of an error and the unacceptable dissemination of accusations about a specific individual without adequate evidence. When publishing, it is advisable to use a synthetic plot and clearly indicate its fictitious nature.

Appendix 29. Case Study: Outdated Information on Criminal Prosecution

Simulated situation: The initial online report mentions an investigation, but the charges are later dropped or the court ruling is amended. In some systems, such material may continue to influence the model's responses. The study records the initial and current procedural situations separately and does not equate the press report with an active search warrant, international notice, or court verdict.

Checkpoints: applicability of Article 10 of the GDPR; source and date of initial publication; existence of an official document; quality of identification of the individual; legal status of the training; possibility of correction and deletion; actual behavior of the model; competent defense route. The expert does not confirm the existence of an INTERPOL entry without confirmation from a relevant competent source.

The report only reports confirmed facts, the risk mechanism, restrictions on access to internal data, and measures to prevent further misattribution. A clear process for providing relevant court documents is created for the applicant, but the burden of reconstructing the entire hidden architecture of someone else's model cannot be shifted to them.

Appendix 30. Case Study: Broker Sells an "Anonymous" Set

Simulated situation: a broker transfers to a developer an archive of forum messages with redacted names, but with rare dates, professions, and geographic descriptions. Individual authors can be re-identified using other open sources. The term "anonymous" in the contract does not resolve the legal issue: an analysis of the means of identification actually available to the parties and reasonably expected recipients is required [1][15].

The auditor requests the anonymization scheme, the legal basis for the initial collection, a description of possible external connections, and the results of re-identification attempts with authorized test data. If the evidence is insufficient, the statement "anonymization status not

confirmed" is used, rather than "illegal collection confirmed." The developer makes a separate decision on the admissibility of further use.

A practical measure is to temporarily restrict the use of the disputed corpus, conduct targeted tests, revise the contractual guarantee, modify the collection if necessary, and establish a procedure for handling future complaints from subjects. The goal of this method is not only to identify the violation but also to prevent further unwarranted dissemination of the material.

Appendix 31. Public Transparency Report: Minimum Indicators

A public report should not state a percentage of legal compliance calculated from an arbitrary sum of checkboxes. Preferably, limited verifiable metrics are used, with a clearly defined population, period, and method. Examples include: the proportion of URLs in the registered version of the corpus for which the source is known; the percentage of records with a timestamp; the number of confirmed requests for subject rights resolved in a given period; the average investigation time for a complex request.

For each indicator, the numerator, denominator, excluded elements, assembly process, and constraints are specified. If the extraction involves multiple linking documents, one URL and one person are not interchangeable. An external reader must distinguish the number of processed requests from the number of unique subjects. This methodology protects both the regulator and the public from misleading statistical data.

Indicator	Numerator/denominator	Limitation
Traceability coverage	Objects with correct connection / all objects of the verified version	Does not automatically confirm legality
Filter recall	Identified control cases / all labeled control cases	Depends on the test sample
Rights resolution	Requests completed/accepted for consideration	Does not measure the legal correctness of the refusal
Source timestamp completeness	Sources marked / all verified sources	Does not prove the relevance of the fact

Appendix 32. Draft proposals to EDPB Guidelines 03/2026

Regarding accountability, it is recommended to recommend a versioned inventory of sources, allowing for verification of the path of a specific data category prior to a training run. However, requiring the open publication of all identifying URLs should not be done when this increases the risk to individuals. Differentiated disclosure to the public, oversight bodies, and confidential independent reviewers will allow for a balance of provability and mitigation.

In the legitimate interest section, it would be appropriate to include practical examples of the "mass collection vs. limited subject set" alternatives, as well as a clearer distinction between measuring model performance and the legal necessity test. A positive qualitative experiment does not replace the balancing of rights, but it does provide verifiable data for conducting it. In the absence of the ability to document the choice of alternatives, the decision should not be described as fully motivated.

Regarding transparency, it's important to emphasize that reliance on Article 14(5)(b) does not imply exemption from public information and protective measures. A structure of searchable source

categories, a specified collection period, and a direct channel for data subjects should be recommended. A system for communicating previously expressed prohibitions to the end developer would be useful for data brokers.

Regarding special categories and criminal offense data, a clear workflow is proposed: "exclude or submit for separate legal review," without the false impression that public availability automatically creates an exception to Articles 9/10. These proposals reflect an independent research position and do not constitute accepted standards.

Proposals for the EDPB consultation should include a clear statement of the proposed clarification, the legal issue, the technical reason for the problem, and the limits of applicability. For example, a methodological recommendation could be proposed to document source categories and the statistical effectiveness of filters for broad untargeted collections. It is important, however, not to present such a proposal as an existing mandatory regulation. Similarly, a request to clarify notifications under Article 14 should be consistent with the fact that the article itself sets out conditions and exceptions; recommendations can clarify application, but not rewrite the law.

The second area involves developing formulations for a real response to a subject when an operator has excluded their record from further training but cannot confirm the complete removal of the impact from the available model. It is recommended to differentiate the results across different layers of the technical system and specify verification limitations. The third area involves independently measuring the quality of filters by language and data class. The fourth area involves requirements for the verifiability of brokers' statements without transferring the entire dataset to an audit firm. These proposals are aimed at the practical implementation of GDPR principles and do not entail creating new, independent powers for the non-governmental Observatoire.

Appendix 33. 90-day plan for the implementation of organizational methodology

The first 30 days are devoted to establishing the actual status: a list of project objectives, the roles of controllers and processors, a source inventory, a cross-border access map, and an initial LIA/DPIA assessment. Each unspecified source is assigned a visible status with a temporary restriction decision. The launch of new large-scale collections without a completed verification of prohibited classes is recommended to be suspended by management decision.

During the second 30 days, technical controls are created: an allow/deny registry, a versioned manifest, high-risk data filtering, a robust decision log, a complaints process, and a deletion protocol for all relevant copies. Tests are based on artificial and legally obtained data. Before rolling out to all systems, a false-confidence assessment of the filter's coverage is mandatory.

Between days 61 and 90, the independent group selects representative sources, tests the LIA and dataset lineage, reviews the rights of subjects, compares the public description with the facts, and prepares a limited conclusion on the subject. Unresolved blocking deficiencies and uncertainties are not covered by the phrase "the implementation plan has been completed."

Period	Result	Acceptance criterion
Day 1–30	Map of goals, roles and sources	Unnamed risks are visible in the register
Day 31–60	Filters, logs, rights workflow	Documented tests have been carried out

Period	Result	Acceptance criterion
Day 61–90	Independent sampling and correction plan	Conclusions are limited to verified facts

Appendix 34. Criteria for acceptance of the report by an independent reviewer

A report may be submitted to management if the following is available: a specific object and period of the audit; a map of the standards and status of governing documents; a list of the evidence obtained; test results; a list of uncovered sources; identified deficiencies and corrective measures; and an independent editorial decision. For high-risk personal data, Articles 9 and 10 must be separately reflected, and for indirect data collection, the grounds and limits of Article 14(5)(b) of the GDPR.

Blocking conditions for acceptance: the absence of any specific processing purpose while mass collection is ongoing; the inability to establish one's own role and counterparties; proven acquisition of special categories without the necessary legal verification; the absence of a mechanism to record a disputed source or consider a subject's request within the available data; and a public declaration of complete legality without relevant evidence. These are proposed organizational quality criteria, not a closed list of violations.

The conclusion should contain an explicit "not verified" statement when access to the internal model, part of the corpus, or historical logs was not provided. Such transparency increases the value of the research: interested parties will be able to distinguish between verifiable facts and hypotheses and raise specific questions within the proper framework.

Appendix 35. Register of legal and technical issues for subsequent review

Before the official publication of the full study, a legal reviewer verifies the wording of the GDPR and the AI Act, the status of EDPB Recommendation 03/2026, the applicability of French provisions and national implementations of the DSM/TDM, and the correctness of special personal data regimes. A technical reviewer verifies the lineage architecture, the distinction between file deletion and weight changes, the reproducibility of the sample audit, and the validity of the proposed uncertainty formula.

It is also necessary to ensure that none of the provided case studies can be interpreted as describing a specific client, nor can any methodological proposal be interpreted as official authorization already received from Observatoire ARGA. Published contact information, logo, and details of the organization must correspond to the actual approved version. If the document is used to prepare a position on an EDPB consultation, it does not replace a self-submitted application and confirmation of submission.

Subject of review	Who checks?	What do you need to get?
Legal relevance	Specialized lawyer	Dated conclusion
Link accuracy	Editor	Active primary sources
Digital methodology	Technical specialist	Testing protocols and restrictions
Statistical examples	Analyst	Recalculated indicators

Subject of review	Who checks?	What do you need to get?
Independence and publication	Editorial organ	Decision, conflicts and scope

Appendix 36. Designing a crawler as a legally controlled system

When designing an automated crawler, it's advisable to separate the policy plane and execution plane. The former contains the approved target, permitted domains, content categories, time windows, regional restrictions, exceptions for vulnerable groups, and maximum retention periods. The latter executes the network of requests according to a specific policy version. Each crawl job must reference an immutable policy ID; otherwise, it's impossible to reproduce the rules the crawler complied with at the time of the disputed operation. The presence of a policy file without mandatory pre-query validation does not constitute effective control.

The semantics of the "allowed domain" rule are particularly important. Subdomains, mirrors, archives, URL parameters, and URL shorteners can lead to unexpectedly expanded approved categories. The project requires URL normalization, domain canonicalization, and mandatory post-redirect validation. For dynamic sites with browser rendering, it is necessary to consider the availability of additional elements not visible in the original HTML. Any change in the way a page is opened can also alter legally relevant access facts, such as authentication or access to a previously hidden section of the page.

Before extraction, a series of checks is performed: resource belonging to an approved category; absence from the deny registry; consideration of explicit technical limitations; valid document type control; exclusion of paths with a high probability of containing sensitive information; limits on the frequency and depth of crawling. If the result is negative, the task must be blocked, and the reason saved as an audit record. The engineer must not be able to surreptitiously disable all fuses to increase the volume of the assembled enclosure.

In a disputed case, a distinction must be made between the absence of a prohibition and the presence of legal authorization. If robots.txt does not contain a prohibition, this does not mean that the processing of the relevant personal data is justified under Article 6 of the GDPR. Conversely, a website's technical prohibition may relate to a contractual or copyright regime, which is being examined in parallel with the GDPR. The EDPB's regulatory draft and the CNIL's practice include the expressed resistance of a resource as a significant assessment criterion; the individual applicability of the rules must be determined on the basis of the facts [2][6].

An independent test is conducted with simulated resources: an allowed domain, a denied path, a redirect to a denied path, an attachment containing personal information, no connection to the rules registry, and a change to the deny registry during execution. The report includes not only the success of blockings but also which steps were unavailable during the audit, whether there were hidden administrative exceptions, who was able to change the rules, and whether their history was saved. The recommended architecture ensures verifiable enforcement of specific restrictions rather than declarative "crawler ethics."

Control object	Failure test	Proof of the result
Policy ID	Task without an approved policy	Launch rejected and recorded
Domain	Redirect to a prohibited resource	Rechecking the final address

Control object	Failure test	Proof of the result
Sensitive path	Link to children's/medical materials	No extraction, there is a ban log
Changing the ban	New deny entry during job	The date and scope of action have been determined
Administrative bypass	Attempting to disable filter without permission	Refusal/separate approval

A crawler is designed with two independent circuits: the access rules and the actual executor. The first approves the target and source classes, prohibited domains, retention periods, maximum volumes, and mandatory actions in case of doubt. The second executes the tasks, but cannot arbitrarily expand the permitted geography or data type simply because it has found new links. A change log is maintained between them: who proposed the rule, who approved it, which version took effect, and which tasks were launched before it was revoked. This allows for demonstrable verification of compliance and promptly halts the collection of a disputed category.

Control tests should cover more than just routine operation. For example, behavior should be checked for unexpected logins, redirects to a medical website, robots.txt changes, the appearance of an API containing data belonging to minors, filter failures, and restarts after failures. In the event of a technical failure, a safe behavior is preferred: temporary suspension or placing the task in a parsing queue, rather than expanding the default collection. Suspension events are recorded as evidence of actual control. The result of such testing does not automatically confirm the legitimacy of the chosen target, so the legal and technical sections of the auditor are kept separate.

Appendix 37. Judicial and evidentiary assessment of the origin of the disputed fragment

In legal disputes, it is often necessary to establish when certain material was published, when it was retrieved, who had control over the content, and in what form it was submitted for training. A simple link to the current web page is insufficient: the page may have changed or disappeared. A version control protocol must exist, taking into account the legality of the storage itself, the hashing of the copy, the network log, and a description of the access context. The disputing party must provide sufficient data to verify the origin without indiscriminately disseminating another's private information.

Timestamps must refer to different events: creation of the original text, publication on the website, first crawler observation, actual download, initial processing, inclusion in the dataset version, and use in the training run. They are not interchangeable. For example, an archived page with a publication date of 2018, first downloaded in 2026, does not prove that the developer had access to the information since 2018. When constructing a timeline, the auditor separately considers possible clock mismatches and alternative sources of timestamps.

Causal relationships between a training document and a specific output of a generative model are particularly challenging. Even a complete phrase match does not preclude the possibility that the text could have originated independently from multiple sources. Without access to training runs, external testing allows one to establish the reproducibility of the output under specified conditions, but does not always establish a specific source. Legal opinions should use separate levels of formulation: fact of extraction; fact of inclusion in the dataset; fact of use in training; fact of reproduction; and the degree of proof of the relationship.

The procedure takes into account the protection of the developer's commercial secrets and the personal rights of others. Limited provision of a technical manifest to an independent expert, anonymized statistics of sources, or judicial access to classified evidence are possible. However, an unreported log cannot be described as verified. The participant is given the opportunity to challenge the identification of the record, the conditions of access, and the time-based statements, and the auditor must preserve evidence of an alternative version.

If an error is detected, it is recommended to create a separate recovery matrix: what is known about the pages, what is known about the copies, what changes are confirmed, and what impact cannot be determined. Such a matrix is especially important for international investigations, sanctions, or financial decisions: the conclusion about the legal status of an individual must be based on a competent official source, not on a match between a search result and a model phrase.

Evidence question	Possible object	Acceptable output
Was there a text?	Versioned copy and attribution	Historical state of the source
Was it loaded?	Crawl log, manifest, hash	The fact of specific extraction
Have you entered the training?	Training run manifest	Use in the agreed version
Was it reproduced?	Repeatable test transcript	The output of the model for this test
What is the causal relationship?	The totality of available facts	Level of evidence with reservations

Appendix 38. Depersonalization, linking, and the risk of reverse attribution

One of the weaknesses of simplified data evaluation is that the identifier filter is tested on individual rows, whereas re-identification can occur through relationships between multiple fragments. A professional position, approximate place of residence, an unusual date, and the description of a rare event can all combine to identify an individual without a direct name. When assessing anonymization, one must consider not only the uniqueness of an individual attribute, but also the possibility of singling out, linkability, and inference within a specific threat model [1][15].

An example of a synthetic test: one training fragment describes a rare job title, a second describes a move to a small town, and a third describes a specific professional award. Individually, the information doesn't identify a person, but when combined with an accessible directory, it can significantly increase the likelihood of identification. The test is conducted on artificial profiles to avoid introducing new risks, and captures a set of external data available to the adversary. The result isn't simply a "name match found"; what matters is the likelihood of linking and the quality of alternative explanations.

In the architecture of large generative models, linking occurs not only within the training corpus. The retrieval layer, external search plug-ins, user queries, and other databases can provide missing attributes. Therefore, the audit at least considers typical downstream scenarios in which the system connects reproducible fragments with new sources. The presence of safe initial cleansing does not guarantee the absence of combinational risks in the finished product.

The evidence form contains a list of verified threats, available auxiliary sets, access rights, mitigation measures, linking resistance, and conditions under which the conclusion requires revision. If real external databases cannot be verified, a limited scope of the conclusion is

established. For confidential forum and medical records, a more stringent exclusion strategy is recommended rather than endless refinement of the probabilistic filter.

Legally, it is necessary to specifically determine for whom and under what conditions the information can be used to identify an individual. The term "pseudonymized" does not exclude the GDPR, and the preliminary assessment of anonymity must be reconsidered when new means of aggregation emerge or accessibility changes. The controller documents the basis for its conclusion, not the degree of its confidence in the technology's designation.

Examination	Scenario	Reporting criteria
Singling out	Rare field combinations	Is it possible to select a record without a name?
Linking	Two approved test bases	Is there a consistent match?
Inference	Derivation of sensitive property	Is it possible to obtain a personal attribute in a reasonable manner?
Model output	Relationship between retrieval and response	Is there still a risk after the filter?
Revision	New sources or access rights	Is a repeat DPIA necessary?

Appendix 39. Managing data deletion from backups and training epochs

Removing a disputed record is a lifecycle management task, not a single file call. A large corpus may have raw snapshots, deduplicated snapshots, tokenized caches, shards, intermediate checkpoints, embeddings, and backup versions. Some technical copies may have an indirect link to the original URL. Therefore, restoring affected objects requires traceability back to the corresponding version build operation; the lack of such ability increases reporting uncertainty [1] [10].

The procedure begins with restricting further access and freezing new learning on the disputed fragment. Associated identifiers, the purpose, and the basis for retaining each copy are then determined. Applicable preservation obligations for certain evidence or backup elements are considered separately; deletion must not destroy legally required information without legal review. The developer documents what exactly is being deleted, what is legally retained, and for how long.

If the model has already been trained on a fragment, technical options include data-free retraining, a validated machine unlearning method, constraint retrieval, and appropriate response filters. No method is declared universally equivalent to erasure. Tests may include canary extraction of artificial sequences, memorization stress tests, and behavioral modification testing on legitimate tasks. The external report specifies the configuration and constraint, but does not claim to prove the absence of any hidden influence.

The system process includes a request log, separate statuses for raw removal, index removal, training future exclusion, and model risk mitigation. Each status has a verification artifact and an execution owner. If a lack of traceability prevents the implementation of certain measures, this must be clearly communicated in accordance with legal requirements, with a realistic plan to address the technical limitation. False confirmation of complete removal could worsen the applicant's position.

For international organizations, contractors' copies should be considered. Changing the name of a corpus or moving it to another jurisdiction does not eliminate the original objective of subject rights. Agreements with associated processors must ensure the coordinated transfer of requests and proof of compliance. The auditor reviews a sample of completed cases, comparing promised actions with actual changes to technical objects.

Object	Action	Completion condition
Raw and clean sets	Exclusion/restriction of access	Agreed manifest and logs
Temporary caches	Removal after the technical deadline	Lifecycle/retention check
Retrieval index	Deletion and reindexing	Search results test
Future training	Introducing negative constraints	Next Run Control
Model weights	Remediation and residual risk test	Limited withdrawal without false guarantee

Appendix 40. Procedure for challenging the decision of an independent auditor

The independence of an external report does not absolve its author from procedural fairness. The organization being audited must be provided with specific established facts, a description of the tests, and a reasonable opportunity to correct inaccuracies before publication, as long as doing so does not destroy evidence or pose a threat to the applicant. However, the right to provide an explanation does not entitle the organization to block a bona fide publication solely because it disagrees with an unfavorable conclusion.

The procedure involves four stages. First, an evidence list is compiled, indicating the availability of materials. Then, preliminary conclusions are formulated, strictly separating observation and evaluation. After receiving the position of the audited party, the auditor records which facts are accepted, refuted, remain disputed, or are inaccessible for independent verification. Finally, the editorial board makes a reasoned decision, assessing the need to protect personal information and sources.

When a dispute involves a specific individual, their rights and safety may require limited disclosure of materials to the other party. For example, publishing the full content of a health complaint simply to illustrate a model error is prohibited. If necessary, a secure, summarized version may be permitted, with identifiers removed and documentary evidence highlighted to refute the disputed claim. It is advisable to establish rules for protecting whistleblowers and confidential sources in advance.

The auditor's role must remain separate from legal representation of any party if this creates a significant conflict of interest. Observatoire ARGA may simultaneously develop the methodology and conduct other activities within its mandate, but any specific conclusion requires disclosure of the connection to funding, a commercial commission, or an interested party. In cases of significant conflict, an external review or recusal is proposed.

The final report contains a dissent or response section if the objection is essential to understanding the evidence. After publication, a procedure for correcting factual errors is in place, with a version history maintained to ensure that a conscientious correction does not turn into an unnoticeable

rewrite of an already circulated text. This increases the reproducibility of the study and reduces the risk of further harm from incorrect peer-reviewed publication.

Stage	Document	Procedural guarantee
Preliminary observations	Evidence list	Fact-checking, not accusation
Organization's response	Response record	Opportunity to present counter-evidence
Editorial decision	Decision memo	Independence and reasons for change
Publication	Redacted report	Protection of personal information
Correction	Version history	Transparent explanation of the edit

Appendix 41. Designing a Clear Scraping Notification: A Test with a Real User

The principle of transparency cannot be achieved by a single, legally sound but illegible document. When collecting publicly available sources, a person may not even know that their publications were used to develop AI, much less the name of the specific model, developer, or intermediary. Article 14 of the GDPR requires notification to the prescribed extent and, where appropriate, exemptions from individual notification under Article 14(5)(b) provide for appropriate alternative measures, including publicly available information [1][2].

For independent verification, we propose using the "five questions" test. Can the reader determine in a few minutes: which organization acts as the controller; for what specific purpose is their information collected; what types of web sources are affected; what rights are available in their situation; and how to contact them and receive a reasoned response? This test is a methodological tool and does not replace the formal requirements of the regulations. The page can include a search by website category and time period, but does not necessarily disclose URLs containing sensitive information.

The notification system should be multi-layered. A brief public page explains the general concept, a second-level document contains mandatory information, and a technical appendix outlines source classes and processing boundaries. A closed set with more precise lists can be prepared for the regulator if disclosure is permissible. When compiling language versions, the consistency of substantive legal statements and the correctness of the concept of "objection" should be verified, rather than relying solely on machine translation of buttons.

An independent user test with different types of affected individuals is useful: the author of a public article, a former forum user, an organization representative, and someone mistakenly associated with a sanctioned entity. Participants are asked to determine where to file a complaint, how to learn about potential processing, and where to find a list of sources. The interview should not require disclosure of actual sensitive information; mock scenarios are used. Success, search time, misconceptions, and the need for additional assistance are measured.

The report should clearly disclose the known completeness of the source catalog. If a company doesn't know the exact contents of the acquired dataset, the statement "you can check all your data in the catalog" is misleading. In such a case, the notification should clearly describe the control limitations and the actual investigation channel, and the developer should have a plan for restoring the origin and processing conditions.

The question of man	Expected interface element
"How could you get my text?"	Source categories and collection period
"Who's in charge?"	Controller name and contact
"What exactly happens to the data?"	Description of collection, training, issuance
"Can I object?"	Work uniform with legal qualifications
"What changes after conversion?"	Separating deletion from a set and measures for a model

Appendix 42. Distinction between private scraping and law enforcement access to data

EDPB Guidelines 03/2026 are limited to collections from external internet sources in connection with generative AI performed by the private sector [2]. This limitation has practical implications: research into law enforcement databases, state access to criminal procedure data, and automated decisions by competent authorities may be subject to other specific rules, including legislation on data processing for law enforcement purposes. All EDPB private-law examples cannot be transferred to such regimes without their own qualification.

ARGA considers the scenario where public information about a criminal case is submitted to a commercial database, and a downstream organization uses the model to screen a client, asylum applicant, or litigant. Private collection and official law enforcement may be distinct processes with their own legal basis, purpose, and competent oversight. If the model produces an erroneous biographical association, correcting the training database does not replace reviewing the administrative or judicial decision.

The information chain must preserve the source of the legal assertion: an official court decision, a press release, a news story, an aggregator, or a generated output. For each stage, the level of reliability, validity period, and refutation capability are specified. If a person's legal status depends on a specific official decision, the media text does not replace the original document. Whether a record has been made in international police cooperation systems cannot be determined from the output of a generative model.

The proposed methodology includes a separate "interregime boundary" section. It asks: who uses the result and under what law; what legal force is attributed to the conclusion; is human review provided; and is an individual entitled to correct the original data and appeal the subsequent decision. This approach avoids reducing all consequences to a single GDPR review, ignoring industry-specific procedural safeguards.

An independent study without access to government databases should not make claims about their contents. It can evaluate publicly available model output, the interface's legal architecture, and documents voluntarily provided by interested parties, but it must clearly define the boundaries of its competence. Properly defining the boundaries enhances the evidentiary value of the work and protects the subjects from the re-dissemination of controversial accusations.

Stage	What can't be replaced	Due diligence
Public news	Official status of the case	Primary competent act
Commercial model	Police base	Limited test of the model

Stage	What can't be replaced	Due diligence
Private scraping	Enforcement fee	Separate mode map
Technical fix	Judicial appeal	Independent procedural path

Appendix 43. Pre-training, retraining, retrieval and processing of user queries

The designation "model trained on open data" conceals several distinct stages. Pre-training can involve large, heterogeneous corpora; fine-tuning typically uses smaller target collections; retrieval-augmented generation accesses an external index without the need to change the underlying model parameters; exploitation stores or processes user queries and logs. Each process has its own goal, supervisor, source, period, and possible legal basis [1][9][10].

The problem is particularly noticeable when executing a deletion query. If a subject's name is retrieved from a retrieval index, deleting or restricting the index can quickly and verifiably change the result. If the information was learned in the model parameters during pretraining, other technical measures are required. If it arose solely from the user's current query, the validity of the initial scraping may not be a primary concern. The procedure should not confuse all three cases.

For analysis, we propose dividing the model map into four levels. Level 1 — original stored data and access to it; Level 2 — prepared training/fine-tuning sets; Level 3 — parameters and checkpoints; Level 4 — service indexes, plugins, prompts, and operation logs. For each, the personal data categories, technical searchability, storage period, and responsible controller are documented. Having a single, general "AI data" register is too crude for reproducible verification.

Continuous learning and subsequent user training are given special consideration. If a commercial system retains prompts for model improvement, the purpose of such use must be separately explained, and the subject's rights and applicable restrictions must be assessed. The basis for responding to a user request for continuous learning cannot be automatically extended to all entered personal data. The ICO rightly highlights the separation of model development goals from downstream applications as an important element of purpose limitation [9].

An independent assessment requires an actual architectural diagram and testability at each level. Without such a distinction, the organization risks responding to the applicant about "data in the model" when only the API log information has been verified. The assessor is obligated to identify this substitution as a limitation, rather than inferring an unknown internal architecture.

Stage	Object of control	Typical fix
Pre-training	Large training set and weights	Exception in next runs, check remainder
Fine-tuning	Specialized records	Rebuild/cancel version if necessary
Retrieval	External indexed database	Deleting a record and reindexing
Operation	Prompts, logs, responses	Storage Policy and Access Rights

Appendix 44. Method of Responding to Common Unproven Legal Assertions

The assertion that "all information online is in the public domain" confuses four different concepts: open technical access, personal data protection, the copyright status of a work, and the conditions for commercial reuse. A legal response requires a clear determination of the legal object to which the phrase refers. Otherwise, the dispute devolves into an exchange of slogans and prevents the assessment of specific processing [1][18].

The assertion "we act in the name of technological progress, so there is a legitimate interest" does not complete the test of Article 6(1)(f). While the interest of innovation or commercial activity may be legitimate, necessity and a balancing of rights must be demonstrated separately. The existence of a positive public purpose does not obviate the obligation to consider the scale of invisible collection, historical data, vulnerable groups, information, and the right to object [2][3][5].

The claim "we completely anonymized the data because we removed names" requires technical verification. Rare combinations of dates and circumstances, geographic and social characteristics, embeddings, or the possibility of linking to another set are preserved. Anonymity must be assessed for the context in question, not inferred from a single transformation. If it is impossible to verify the presented means, it is permissible to state that the claimed property is not supported by available evidence [15].

The statement "data removed from the model" requires a definition of the object: raw data, clean dataset, retrieval index, fine-tuning checkpoint, weights, or output filter. Each requires confirmation. It is impossible to promise complete removal of influence without a testing methodology; nor can it be claimed that every fix is impossible. The developer must provide an informed and technically honest answer to a specific request [10].

Finally, the assertion "we have an external certificate, so the collection is legal" displaces the scope of verification. Certification of a management system, such as security, can confirm certain organizational mechanisms, but not the legality of each source or basis for processing. Public verification must specify the subject of the certification, the applicable standard, and the limitations. This standard of response allows Observatoire ARGA to avoid taking a preemptive position of accusation or justification.

Statement	What kind of verification is required?
"Everything is open and permitted"	Right of access, GDPR, content, contracts
"Innovation is sufficient interest"	Three stages of LIA
"Names removed - data anonymized"	Linking/inference and identification context
"Removed from AI"	Model level and proven effect
"The certificate closes the questions"	Scope of certification and subject of the standard

Appendix 45. Case study of negotiations with a broker: a detailed procedure for evidence-based procurement

Suppose a developer wishes to acquire a large archive of publicly accessible messages, but the broker refuses to disclose the initial list of sites, citing commercial confidentiality. The buyer

initially requests not the entire unprocessed corpus, but a categorical manifest: countries and domain types, periods, stated basis for collection, exclusions for personal stories, indicators of 9/10 articles, content rights, and a subsequent deletion mechanism. This request combines the need for verification with the principle of minimizing data disclosure [1][2].

The next step is selective verification. The parties can agree to verify a limited number of records, independently selected, in a closed environment, with the sampling direction, known population, and confidentiality method specified. Trade secrets should not be reduced to the empty statement "the broker's lawyer saw it, so it's all legal." However, the auditor has no right to demand disclosure of sensitive information beyond the permissible scope simply for the sake of analysis convenience.

Following preliminary testing, the agreement should define prohibited classes of sources, the obligation to report any disputed data discovered, a process for jointly fulfilling requests from subjects, and the ability to promptly exclude entire subsets before the investigation is completed. Technical version identifiers, a logic for prohibiting training on old batches, and conditions for suspending deliveries in the event of a significant refusal to provide evidence are established.

Applicable law and jurisdiction are considered separately from the source. Even if the supplier is based outside the EU, GDPR obligations may arise if the relevant conditions are met. However, the purchase of a data set by a European organization does not automatically mean that all data pertains to EU residents. It is necessary to analyze territorial factors, civil rights, and privacy rights based on the actual nature of the processing.

The final decision is categorized as "accept subject to conditions," "request additional data," "exclude specific categories," "postpone procurement," and "reject." Each decision is accompanied by written reasons. If a material class of data cannot be verified, formal acceptance of the risk by management does not negate the obligation to comply with the law; it merely reflects organizational responsibility for the chosen position.

Procurement stage	Document	Blocking circumstance
Preliminary assessment	Purpose/legal basis map	The controller is not defined
Origin verification	Category manifest	There is no traceability
Selective viewing	Restricted Access Protocol	Uncontrolled disclosure of PII
Agreement	Rights/remediation clauses	Unable to fulfill subject's request
Start of training	Dataset version permission	Unclosed high-risk categories

When pre-approving the procurement of a corpus, the lawyer, data engineer, and data protection officer should work on a single version card. The lawyer establishes applicable bases and limitations, the data protection officer verifies the actual structure and reproducibility of the filters, and the data protection officer assesses the risks to people. If the procurement manager receives only three unrelated approvals, it is unclear whether the findings belong to the same set. Therefore, an end-to-end delivery identifier and a blocking rule are necessary: a significant change in source categories or processing method triggers a re-evaluation of the card.

The contract may establish penalties for false representations, but these do not replace the obligation to prevent unlawful processing. Particular attention is paid to the terms of contract termination: retention of backup copies, return of original data, assistance with subject requests,

deletion of working copies from subcontractors, and exclusion of material from future updates. If the seller does not control the versions already resold, it is impossible to promise global deletion to all affected parties. The final audit report must separately contain confirmed facts, unresolved restrictions, and a specific decision: approval, limited approval, deferment, or refusal.

Appendix 46. Justification for disproportionate effort in individual notification

The exception provided for in Article 14(5)(b) of the GDPR should not automatically result in a waiver of the transparency obligations. To apply this exception, the controller must examine the impossibility or disproportionate nature of individual notification, taking into account the characteristics of the data and the impact of the lack of notification on the data subjects. EDPB draft regulation 03/2026 discusses the number of data subjects, the age of the information, and possible safeguards; it emphasizes the need to maintain public access to the information provided [1][2].

The methodology proposes justifying disproportionality using five questions. First, does the controller have a reliable means of contacting affected individuals without additional bulk processing? Second, what is the estimated coverage of individuals and the degree of duplication? Third, what actions would need to be taken simply for notification, and what new risks would they create? Fourth, to what extent would the lack of notification hinder an individual's ability to exercise their rights? Fifth, what public, technical, and organizational measures effectively mitigate these consequences?

Two case studies demonstrate the limits of generalizations. In the first, the project processes several hundred voluntarily submitted professional resumes with valid email addresses, and the mass assignment of tasks alone does not render individual notification disproportionate. In the second, a historical library of forum discussions is processed; many posts are pseudonymous, contact information is missing, and individual messages would require re-matching with social media. Here, the circumstances may support the argument that notification is excessive. However, a final legal conclusion requires a full analysis, not a simple cost arithmetic.

If the controller invokes an exception, the public notice must contain the essential information required by Article 14 and a readily accessible method of appeal. Simply posting the generic phrase "we may use data from the internet" on the website is not sufficient. The categories of sources, the purpose, the grounds for processing, the timeframes, and the responsible parties must be defined. If there are gaps, the boundaries of the register must be clearly communicated, rather than an inventive approach.

The independent review includes a random check of available contacts, an internal cost assessment without access to unnecessary data, and a user accessibility test of the public notice. The report should clearly outline the facts, the legal assessment, and the extent to which the result depends on the assumptions. If conditions change over time, the decision is subject to review.

Factor	What to confirm	Possible objection
Number of subjects	Realistic evaluation method	Duplicate messages
Address availability	Possibility of communication without new collection	Incompleteness and obsolescence
Data age	Distribution by periods	New information may be more accessible

Factor	What to confirm	Possible objection
The harm of failure to notify	Consequences for the implementation of rights	Insufficient alternative measures
Public information	Accessibility and completeness	A formal page without meaning

Appendix 47. National Legal Restrictions on Public Registers: Conflict Map

A public registry does not represent a uniform legal regime. Differences exist between commercial, judicial, professional, administrative, and archival systems; access may be for a specific public purpose and subject to restrictions on bulk downloading, reuse, individual search, or historical preservation. The developer should establish the conditions for each class of resource and the relevant jurisdiction, rather than generalize from the fact that a web interface is accessible [1][2].

When studying various countries, it is necessary to distinguish four layers: rules for access to the primary registry; grounds for processing personal data during independent extraction; the permissibility of further commercial or research processing; and rules for using the results in decisions about a specific individual. Lawful access to a public record may not provide sufficient grounds for creating a global biographical risk database. In some contexts, restrictions on the preservation of information about offenses and court decisions are of additional importance.

A typical technical risk arises when integrating multiple registries, when identical names and similar dates create erroneous personal associations. Therefore, regardless of access conditions, rules for identification, attribution of primary legal status, time stamping, and result review are required. Erroneous associations with criminal prosecutions, sanctions, or insolvency should not become independent, machine-generated legal facts.

The jurisdictional map includes the name of the country and authority, the source category, the public purpose, the type of potential personal data, the legal status of mass extraction, the presence of special rules for reuse, and possible cross-border access. National legal sources should be reviewed by a specialist lawyer for a specific project: this international report does not purport to be an exhaustive list of current prohibitions in all countries.

It's useful for independent auditors to apply the "no carryover" rule: permitting one category of registers in one country does not imply the permissibility of a similar collection from another country. The organizational decision is accompanied by a date and a reference to the applicable act or official policy, and in the absence of reliable material, a status pending legal review is established. This model reduces the risk of surreptitious substitution of a convenient general slogan for international comparative work.

Analysis layer	Security Question
Primary access	Who publishes the register and under what conditions?
Mass gathering	Is automatic transaction volume allowed?
Personal data	Which GDPR/national regulations apply?
Information from Article 10	Is there a special legal permit?
Downstream solution	How is the current legal status verified?

Appendix 48. Statistical comparison of the quality of two filtration systems

Comparisons of filter versions should be based on a fixed benchmark set, clear labeling, and predefined metrics. If the team changes the test set simultaneously with the algorithm, the observed improvement may be explained by a reduction in test complexity. Independent labeling, temporary fixation of criteria, and a separate adversarial set for rare cases are necessary. Results for one language do not automatically transfer to others.

Let's assume that Filter A is tested on 100 synthetic examples with special-category data and detects 82, while Filter B on the same sample detects 92. On this set, the observed sensitivity has increased from 82% to 92%, or by 10 percentage points. However, we need to keep a look-through of the paired data: which specific 10 examples were detected and whether new missing data emerged for other types. A simple comparison of two percentages does not establish the statistical significance of the change or characterize false positives.

For legal purposes, the cost of each error is important. If a false negative results in health data being included in the training set, the consequences can be more serious than a simple deterioration in the dataset's quality. On the other hand, an excessive filter can remove legitimate data from a particular language, culture, or scientific discipline, thereby distorting the model's results. Selecting a threshold requires an analysis of both sides of the risk without the need for legally fictitious "legality scores."

When subsequently using OCR, translation, or segmentation, the context changes: the filter must be tested not only on the original strings but also on real transformations. The report describes the known conditions of each test, the statistical uncertainty, and scenarios not covered by the test. The test environment should not contain real, redundant personal data; if legally obtained real materials are used, their processing is separately justified.

An independent expert can repeat the calculations on the provided confusion matrix and verify the compliance of the selected control set with the stated use. If only final marketing figures are provided without the original calculations, the permissible conclusion is limited to a statement of incomplete evidence. Comparison of filter quality is useful as an engineering control, but does not replace verification of the legal basis and purpose of these operations.

Indicator	Formula	What it doesn't prove
Recall	TP/(TP+FN)	Legality of sources
Precision	TP/(TP+FP)	Completeness of detection of rare cases
False negative rate	FN/(TP+FN)	No errors outside the test
The difference between recall	Recall B – recall A	Causality and significance in themselves
Coverage of sources	Tested objects/target population	Full GDPR compliance

Appendix 49. Incident during the opening of the training corpus: a combination of two procedures

A separate issue arises when personal data is unlawfully transferred to a training corpus, and then the storage facility itself is accessed by an unauthorized party. These events cannot be lumped

together. The first concerns, in particular, the lawfulness, minimization, and transparency of processing; the second may constitute a breach of personal data security within the meaning of Article 4(12) of the GDPR and, under appropriate circumstances, trigger an assessment of the notification obligation under Articles 33 and 34 [1].

A practical scenario: a contractor publishes a diagnostic file containing several thousand records containing names and sensitive personal data. The organization must immediately secure the infrastructure, preserve evidence of the event, determine the categories and number of possible actors, the nature of the compromise, and the potential consequences. At the same time, it is necessary to determine why this data was in the training pipeline in the first place, whether there was a legitimate reason for its extraction, and why filters failed to remove unnecessary fields.

A separate legal assessment is essential for timelines. Notifying a personal data breach to a supervisory authority under Article 33 of the GDPR has its own risk criteria and timing, while a public dispute over the legality of scraping does not automatically trigger the same timeline. If there is a risk to rights and freedoms, an analysis is conducted of the need to notify affected individuals under Article 34 and the applicable exceptions. The report should include the date of awareness, the recipients, the decision, and the supporting evidence, rather than automatically applying a single internal template to all incidents.

Remediation requires eliminating access, verifying linked copies, assessing the presence of fragments in the current model, and modifying the process design that created the risk. If an organization merely revokes the public link to the file without reviewing the unjustified storage or training data lineage, the root cause may persist. A separate assessment of the systemic change after the operational threat has been resolved is recommended.

An independent audit must maintain the confidentiality of evidence, not publish the compromised data itself, and not expand the number of individuals who have gained access. The conclusion must formulate separately the following: the identified technical incident, the legality of the original processing, and the quality of the remediation. This prevents a situation where a government reporting issue conceals a separate structural flaw in source management.

Direction	Item	Proof
Security incident	Unauthorized access	Logs, timeline, scope
Initial collection	Legality of hitting the body	LIA, source register
Storage	Minimization and timing	Retention records
Notifications	Risk, recipients and timing	Decisions under Articles 33/34
System fix	Preventing recurrence	Pipeline and access inspection

Appendix 50. Educational file: request for deletion due to name overlap

A generative model generates a claim about an alleged criminal case against a citizen with a common first and last name. The applicant submits a request to have the information removed. An old report about another person with the same first and last name does indeed exist online. When reviewing the case, the simple logic of "the record was found, therefore it's about the applicant" cannot be accepted, nor can the investigation be completely dismissed due to the complexity of identification.

The first step is to securely determine what exactly was generated, in what model configuration, and under what parameters. The minimum fragment necessary for evaluation is preserved, without disseminating the full personal profile. The second step is to separate the possible primary source from the individual's legal classification. If the information relates to criminal offenses, Article 10 of the GDPR and applicable additional provisions are analyzed. The third step is to identify possible internal training/retrieval links, if access is permitted and technically feasible.

To verify identity, an organization must not collect unnecessary documents from the applicant or disclose information about another person. A strictly limited check of individual attributes may be required, but it must be proportionate and documented. If a match is confirmed as erroneous, the decision to correct includes removing the relevant link from available indexes, prohibiting re-collection of the disputed record in the applicant's context, and measures to reduce the recurrence of the incorrect finding.

A separate legal issue is the potential processing of data belonging to another, genuinely named individual. Correcting one false attribution does not automatically legitimize training on the original criminal material. An independent assessment is conducted for the original document and each subsequent processing. This issue cannot be resolved by rearranging the names in a publicly available response.

The appeal is concluded with a reasoned letter outlining the actions taken and their limitations. If only a change to the output filter has been established, do not write "we completely removed the data from the model." If the technical origin remains unknown, the limitation must be communicated in an acceptable manner and the relevant investigation must be continued. This scenario emphasizes that protecting a person's data is a legal process with digital evidence, not a quick search for a similar string.

Complexity	The right measure
Namesake	Do not identify by name
Criminal context	Separate Article 10 of the GDPR
No full origin	To designate the limit of establishment
There is a model issue	Save a limited reproducible test
Only the output has been corrected	Don't promise to change all the scales

When a name matches, it's necessary to distinguish between identity verification and data identification in the information system. A claimant can verify their identity, but this doesn't prove that the disputed record pertains to them. Conversely, a name match in a vector search could return data belonging to someone with the same last name, and a careless response to an access request could reveal sensitive information to a third party. Therefore, the procedure uses sufficient additional criteria only where legal and necessary, and discourages the collection of a complete set of identification documents simply to correct a public fact.

The training file must include four outcomes. First: the data does indeed relate to the applicant, its further processing is deemed inadmissible, and specific corrective measures are confirmed. Second: the record belongs to another person, and the model's inference arose from a false match; the matching procedure is adjusted without revealing the true subject's data. Third: the information was not found in accessible systems, but the absence of sources does not preclude external copies; this is

clearly explained. Fourth: the evidence is insufficient, and limited additional verification is required. Each outcome is accompanied by an indication of what was investigated, which system versions were covered, and whether a resubmission is possible.

Sources and regulatory and methodological materials

As of 22 October 2026. EDPB draft guidance documents are noted separately; the sources cited do not constitute evidence of in-house field verification.

[1] The European Parliament and the Council. Regulation (EU) 2016/679 (GDPR), current text; Articles 5, 6, 9, 10, 12–22, 25, 28, 32, 35, 36 and Chapter V. Official source

[2] EDPB. Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, accepted for public consultation on 7 July 2026 (draft). Official source

[3] EDPB. Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, December 17, 2024. Official source

[4] EDPB. Guidelines 07/2020 on the concepts of controller and processor in the GDPR, final version 7 July 2021. Official source

[5] EDPB. Guidelines 1/2024 on processing of personal data based on Article 6(1)(f) GDPR, consultation version; not yet final. Official source

[6] CNIL. La base légale de l'intérêt légitime: mesures à prendre en cas de collecte des données par web scraping, June 19, 2025. Official source

[7] ICO. The lawful basis for web scraping to train generative AI models, response to the consultation series. Official source

[8] CNIL. AI system development: recommendations to comply with the GDPR, including data collection using web scraping. Official source

[9] ICO. Purpose limitation in the generative AI lifecycle, response to the consultation series. Official source

[10] ICO. Engineering individual rights into generative AI models, response to the consultation series. Official source

[11] The European Parliament and the Council. Regulation (EU) 2024/1689 (AI Act), Article 53 in its consolidated version of 27 July 2026. Official source

[12] European Commission. General-purpose AI obligations under the AI Act; Summary of obligations and public summary of training data. Official source

[13] EDPB. Public consultation: Guidelines 03/2026 on web scraping in the context of generative AI, 8 July – 30 October 2026. Official source

[14] EDPB/WP29. Guidelines on transparency under Regulation 2016/679 (WP260 rev.01), adopted by WP29 and endorsed by the EDPB. Official source

[15] EDPB. Guidelines 02/2026 on Anonymisation, version 1.0, for public consultation (draft). Official source

[16] CNIL. Privacy Impact Assessment (PIA): Methodology and Materials for DPIA. Official source

[17] Court of Justice of the EU. Case C-136/17, GC and Others v CNIL, Grand Chamber judgment of 24 September 2019, full text on the legal limits of de-indexing sensitive data. Official source

[18] European Parliament and the Council. Directive (EU) 2019/790 on copyright law in the Digital Single Market, Articles 3 and 4. Official source

[19] European Parliament and the Council. Regulation (EU) 2024/1689 (AI Act), Article 111(3): transitional arrangements for providers of GPAI models placed on the market before 2 August 2025. Official source