NOTE D'ANALYSE

Cabinet TFEYL — Droit des nouvelles technologies • IA • RGPD • Droit international

Objet	Lignes directrices 03/2026 du Comité européen de la protection des données (EDPB) sur le web scraping dans le contexte de l'IA générative
Référence	EDPB, Guidelines 03/2026 on web scraping in the context of generative AI, version 1.0, adoptées le 7 juillet 2026 pour consultation publique
Destinataires	Clientes et clients du cabinet — directions juridiques, DPO, DSI, directions de l'innovation
Date de la note	28 août 2026
Classification	Note d'analyse — diffusion clients

I. L'essentiel en cinq points

1. Le scraping n'est pas interdit — il est encadré. L'EDPB ne prohibe pas la collecte de données publiquement accessibles pour entraîner des modèles d'IA générative. Il la soumet au droit commun du RGPD, avec une exigence de démonstration : c'est au responsable de traitement de prouver, mesures à l'appui, que sa collecte est licite, nécessaire et proportionnée.

2. L'intérêt légitime devient la voie étroite. Le consentement est jugé pratiquement inapplicable à grande échelle — et l'absence de fichier robots.txt ne vaut pas consentement. Reste l'intérêt légitime (art. 6.1.f), subordonné à un test en trois conditions cumulatives, dont la balance des intérêts ne peut pencher en faveur du responsable qu'au prix de **mesures d'atténuation concrètes et documentées**.

3. La transparence individuelle peut être écartée, mais jamais la transparence tout court. L'impossibilité ou l'effort disproportionné (art. 14.5.b) doit être justifié par une pesée documentée ; l'information publique — politique de confidentialité détaillée, sources, caractéristiques du crawler — demeure **toujours** obligatoire.

4. Les données sensibles : une tolérance résiduelle, sous conditions strictes. S'inspirant de l'arrêt GC & Others (C-136/17), l'EDPB admet que la collecte *incidente et résiduelle* de données de l'article 9 puisse être analysée « dans le cadre des responsabilités, pouvoirs et capacités » du responsable — à condition de filtres préalables, de suppressions immédiates et d'une surveillance continue des sorties du modèle. Ce n'est **pas** une exemption générale.

5. L'accountability est le cœur du dispositif. Cartographie des données, horodatage, validation des sources, registre des critères d'exclusion, analyse d'impact, réévaluation régulière des mesures : la charge de la preuve pèse sur le responsable de traitement. En cas de contrôle, ce qui n'est pas documenté sera réputé ne pas exister.

II. Le contexte : pourquoi ces lignes directrices comptent

Le web scraping est la matière première de l'IA générative : extraction automatisée, à grande échelle, de contenus publiquement accessibles — textes, images, enregistrements — pour constituer les jeux de données d'entraînement. Jusqu'ici, la pratique évoluait dans une zone grise : les données sont « publiques », mais leur réutilisation massive, invisible pour les personnes, heurte frontalement plusieurs principes du RGPD.

Les lignes directrices 03/2026, adoptées le 7 juillet 2026 pour consultation publique, constituent la **première doctrine d'ensemble** du Comité européen de la protection des données sur ce sujet. Elles s'inscrivent dans la continuité de l'avis 28/2024 du 17 décembre 2024 sur les modèles d'IA et des lignes directrices 1/2024 sur l'intérêt légitime, et font écho à la déclaration conjointe du 23 février 2026 sur les images générées par IA. Leur portée est limitée au scraping opéré par des **entités privées** aux fins d'entraînement d'IA générative — qu'il soit réalisé en interne ou sous-traité.

> **Portée juridique.** Des lignes directrices de l'EDPB ne sont pas des actes contraignants. Elles n'en constituent pas moins la doctrine que les autorités nationales de contrôle — la CNIL en tête — appliqueront en cas de contrôle ou de plainte. À ce titre, elles fixent dès maintenant le standard de diligence attendu.

III. Le cadre juridique applicable : ce que l'EDPB retient

3.1 Le champ : quatre opérations, un seul régime

L'EDPB décompose le scraping en quatre étapes — définition des critères de collecte, extraction, nettoyage, structuration et stockage — et constate que chacune constitue un traitement au sens du RGPD dès lors qu'elle porte sur des données à caractère personnel, directement (nom, image, coordonnées) ou indirectement identifiantes. Les jeux de

données « mixtes » — mêlant données personnelles et non personnelles — n'échappent pas au RGPD pour la part personnelle qu'ils contiennent.

La qualification des acteurs — responsable de traitement, responsables conjoints, sous-traitant — relève d'une analyse au cas par cas selon le degré d'implication de chacun dans la constitution du jeu de données. Point de vigilance pour les entreprises qui **achètent** des jeux de données scrapés : l'acquéreur peut être qualifié de responsable de traitement de la phase d'utilisation, voire de responsable conjoint s'il a participé à la détermination des finalités et moyens de la collecte.

3.2 Les principes cardinaux, déclinés au scraping

Limitation des finalités. La finalité doit être déterminée, explicite et légitime *avant* la collecte. Pour les modèles à usage général dont l'usage précis n'est pas encore arrêté, l'EDPB recommande de se référer à l'objectif poursuivi par le développement du modèle — commercial, public, recherche scientifique, usage interne ou externe.

Minimisation. C'est l'exigence la plus opérationnelle du texte. L'EDPB attend des mesures *avant* la collecte — envisager les données synthétiques, définir des critères précis, cartographier, filtrer, exclure les sources qui s'opposent structurellement au scraping (robots.txt, CAPTCHA, ai.txt) — et *après* la collecte : filtrage syntaxique (expressions régulières pour écarter numéros de sécurité sociale, téléphones), anonymisation ou pseudonymisation dès que possible.

Exactitude. Trois recommandations concrètes : scraper des sources fiables et maintenues (registres officiels plutôt qu'agrégateurs), horodater chaque collecte, valider les données avant usage en entraînement. L'exigence s'étend au résultat : le modèle, une fois sur le marché, ne doit pas produire de données personnelles inexactes.

Transparence. L'information individuelle des personnes peut être écartée si elle s'avère impossible ou exige un effort disproportionné (art. 14.5.b) — mais l'EDPB encadre strictement le recours à cette exception : pesée documentée à l'échelle du jeu de données (nombre de personnes, ancienneté des données, garanties adoptées), et avertissement que l'exception ne doit pas être invoquée « de routine » hors des finalités d'archivage, de recherche ou de statistique. Dans tous les cas, une mesure demeure impérative : **rendre l'information publique**, via une politique de confidentialité indiquant catégories de données, finalités, base légale, sources précises — y compris leur caractère public ou non — et caractéristiques du crawler.

3.3 La base légale : l'intérêt légitime, voie étroite

Le consentement est écarté d'emblée : impossible à recueillir à grande échelle auprès de personnes avec lesquelles le responsable n'a aucune relation. L'EDPB est formelle sur un point qui fera jurisprudence pratique :

> « L'absence ou la non-applicabilité d'un fichier robots.txt sur un site web ne vaut pas consentement au sens du RGPD. »

Reste l'intérêt légitime (art. 6.1.f), subordonné à trois conditions **cumulatives** :

Condition	Contenu	Application au scraping IA
1. Intérêt légitime	Licite, clairement articulé, réel et présent (non spéculatif)	Admis pour : agent conversationnel, détection de fraude, détection de menaces ; pour un modèle généraliste, articuler l'objectif poursuivi
2. Nécessité	Le traitement permet-il la finalité ? Existe-t-il une voie moins intrusive et aussi efficace ?	Restreindre les critères de collecte plutôt qu'aspirer tout le web ; envisager données pseudonymisées ou synthétiques
3. Balance des intérêts	Les droits et libertés des personnes ne prévalent-ils pas ?	Appréciée au regard de la nature des données, du contexte, des conséquences, des attentes raisonnables — et des mesures d'atténuation

Sur la balance des intérêts, l'EDPB retient plusieurs enseignements structurants. Plus les données sont sensibles ou intimes, plus leur poids est lourd dans la balance — avec une vigilance particulière pour les données de mineurs. Le volume, le nombre de personnes et le caractère indifférencié de la collecte aggravent l'ingérence. Une conséquence technique majeure est soulignée : **à l'état de l'art, une donnée personnelle apprise par un modèle ne s'en efface pas aisément** — ce qui pèse contre le responsable dans la pesée. Enfin, les attentes raisonnables des personnes ne se déduisent pas de la seule information qui leur est donnée : publier en ligne ne signifie pas s'attendre à être scrapé pour toutes finalités. À titre de règle pratique, plus la donnée a été rendue visible au monde entier par la personne elle-même, plus l'attente d'une réutilisation est plausible ; un billet de blog public n'appelle pas la même analyse qu'une publication sur réseau social à accès restreint.

3.4 Les mesures d'atténuation : la condition de recevabilité de l'intérêt légitime

C'est le passage le plus important du texte pour la pratique. L'EDPB énumère les mesures susceptibles de faire pencher la balance — liste non exhaustive, à apprécier au cas par cas :

Avant et pendant la collecte : exclusion par défaut de certaines catégories de données et de certaines sources ; limitation aux données librement accessibles ; respect des signaux d'opposition (robots.txt, ai.txt, CAPTCHA) ; identification claire du crawler via son user agent ; publication d'une liste actualisée des sites scrapés.

Après la collecte : filtrage syntaxique des données identifiables par leur format ; suppression ou anonymisation des données personnelles inutiles dès que possible ; pseudonymisation (remplacement des identifiants directs par des placeholders pendant l'entraînement).

À l'échelle du modèle : mesures limitant les risques de mémorisation, de régurgitation et d'attaque (déduplication notamment) ; facilitation de l'exercice des droits, y compris un **droit d'opposition préalable discrétionnaire** et des listes d'opt-out permettant à une personne de s'opposer à la collecte *avant* qu'elle n'intervienne.

Deux exemples encadrent la doctrine. Une organisation qui collecte massivement des enregistrements vocaux publics pour commercialiser un outil de génération de voix, **sans aucune mesure de protection**, ne peut se prévaloir de l'intérêt légitime (exemple 6). À l'inverse, une organisation qui développe un système génératif de texte à partir de sources librement accessibles, en excluant les contenus manifestement non destinés à être publics et ceux protégés par le droit d'auteur, avec des garanties contre la mémorisation, la facilitation des droits et une politique de confidentialité précise, peut en principe considérer la balance comme satisfaite (exemple 7).

3.5 Les données sensibles : la tolérance résiduelle de l'arrêt GC & Others

Le principe est l'interdiction (art. 9.1) : tout scraping *intentionnel* de données sensibles exige une dérogation de l'article 9.2, en sus d'une base légale de l'article 6. Mais l'EDPB reconnaît une réalité technique : malgré les mesures de prévention, une collecte **incidente et résiduelle** de données sensibles peut survenir.

S'inspirant de l'arrêt GC & Others (CJUE, 24 sept. 2019, C-136/17) relatif aux moteurs de recherche, le Comité admet que l'interdiction puisse s'appliquer au responsable « dans le cadre de ses responsabilités, pouvoirs et capacités » — sous quatre conditions cumulatives : similarité du traitement avec celui d'un moteur de recherche ; caractère exclusivement incidente et résiduel de la collecte ; difficulté ou impossibilité d'évaluer l'ampleur de la collecte avant qu'elle n'intervienne ; et mise en œuvre effective de mesures à chaque stade — filtres et critères d'exclusion *avant* collecte (notamment des sites structurellement sensibles ou fréquentés par des mineurs), suppression immédiate *après* collecte (y compris sur demande plausible d'une personne), filtres de sortie et tests de résistance aux attaques *pendant* le développement, surveillance continue des productions du système *après*

déploiement, le cas échéant jusqu'au « machine unlearning » lorsque la technique le permettra.

> **Avertissement de méthode.** L'EDPB insiste : cette construction n'est pas une exemption générale aux articles 9 et 10. Elle suppose une analyse au cas par cas, documentée, et une capacité à **démontrer** la pertinence et l'efficacité des mesures — avec réévaluation régulière et veille sur l'état de l'art.

IV. Ce que cela change pour vous

4.1 Vous développez ou entraînez des modèles d'IA

Votre feuille de route de conformité comporte désormais sept chantiers : la définition documentée de la finalité avant toute collecte ; la cartographie des sources avec critères d'exclusion ; le respect des signaux d'opposition technique ; la politique de transparence publique détaillée (sources, crawler, catégories) ; le test d'intérêt légitime écrit en trois conditions ; les mesures anti-mémorisation et les filtres de sortie ; et la procédure de gestion des droits, incluant idéalement un mécanisme d'opt-out préalable. L'analyse d'impact (AIPD) sera dans la plupart des cas indispensable, le scraping massif réunissant plusieurs critères de haut risque.

4.2 Vous achetez des jeux de données ou des modèles pré-entraînés

La doctrine EDPB sur la qualification des acteurs vous expose : l'acquisition de données scrapées par un tiers ne vous exonère pas. Vos contrats doivent désormais comporter des **garanties de provenance et de conformité** des jeux de données (origine des sources, base légale du collecteur, mesures d'atténuation mises en œuvre, existence d'un mécanisme d'opposition), des clauses d'audit et des stipulations de répartition des rôles au sens des articles 26 et 28 du RGPD. À défaut, vous héritez du risque.

4.3 Vous êtes éditeur de site ou de contenus

La doctrine est doublement utile. Elle confirme que robots.txt, ai.txt et CAPTCHA constituent des **signaux d'opposition** que les scrapeurs diligents doivent respecter — leur mise en place est un acte de gouvernance de vos actifs informationnels. Et elle rappelle que l'absence de ces signaux ne vaut pas consentement : votre position juridique en cas de scraping non autorisé s'en trouve confortée, y compris au regard de l'exception « text and data mining » de la directive 2019/790, que l'EDPB cite en filigrane de l'exemple 7.

V. Checklist de conformité — scraping et IA générative

#	Action	Référence	Priorité
1	Documenter la finalité avant toute collecte (y compris pour un modèle généraliste : objectif poursuivi)	Art. 5.1.b	Immédiate
2	Rédiger le test d'intérêt légitime en trois conditions, avec pesée documentée	Art. 6.1.f	Immédiate
3	Définir les critères de collecte et la liste des sources exclues (sites sensibles, sites de mineurs, sources opposées)	Art. 5.1.c	Immédiate
4	Respecter robots.txt, ai.txt, CAPTCHA ; identifier le crawler (user agent)	GL 03/2026 § 65	Immédiate
5	Publier la politique de transparence : catégories, finalités, base légale, sources précises, caractéristiques du crawler	Art. 14	Immédiate
6	Mettre en place le filtrage syntaxique et l'anonymisation/pseudonymisation post-collecte	Art. 5.1.c, 32	30 jours
7	Instaurer un mécanisme d'opposition (opt-out préalable, liste d'exclusion)	Art. 21 ; GL § 65.f	30 jours
8	Déployer les mesures anti-mémorisation et les filtres de sortie ;	GL § 65.i, § 74.iii	60 jours

	tester la résistance aux attaques		
9	Réaliser ou actualiser l'AIPD ; prévoir la réévaluation périodique des mesures et la veille état de l'art	Art. 35 ; GL § 76-77	60 jours
10	Réviser les contrats d'acquisition de données/modèles : garanties de provenance, audit, répartition des rôles	Art. 26, 28	90 jours

VI. Ce qu'il faut retenir

L'EDPB ne dit pas « n'aspirez plus le web ». Il dit : **aspirez moins, aspirez mieux, et prouvez-le**. La conformité d'un programme de scraping ne se mesurera plus à l'intention déclarée, mais à la chaîne documentaire — critères d'exclusion, filtres, horodatage, test de balance, mécanismes d'opposition, surveillance du modèle. Pour les acteurs sérieux, c'est une bonne nouvelle : la doctrine dessine une voie praticable, et l'exemple 7 montre qu'un scraping conforme est possible. Pour les autres, le message est sans ambiguïté : la collecte massive sans garde-fous ne passera plus.

Ces lignes directrices sont adoptées pour **consultation publique** : leur version finale pourra évoluer, mais il serait imprudent d'attendre — l'avis 28/2024 et les lignes directrices 1/2024, déjà en vigueur, en portent l'essentiel.

Le Cabinet TFEYL accompagne ses clients dans la sécurisation de leurs programmes de collecte et d'entraînement IA : audit de conformité scraping, rédaction des tests d'intérêt légitime, AIPD, négociation des garanties contractuelles de provenance des données. Réponse sous 48 heures.

Cabinet TFEYL — 18 rue de la Pépinière, 75008 Paris

+33 6 98 57 35 67 • mtfeyl@tfeyl-avocat.com

Note d'analyse établie sur la base des Guidelines 03/2026 de l'EDPB (version 1.0, 7 juillet 2026, consultation publique), de l'avis EDPB 28/2024 (17 décembre 2024), des lignes directrices 1/2024 sur l'intérêt légitime (8 octobre 2024) et de la jurisprudence CJUE (GC &

Others, C-136/17, 24 septembre 2019). Document d'information juridique générale ; il ne se substitue pas à une consultation personnalisée. © 2026 Cabinet TFEYL. Tous droits réservés.