JurisPONt
Analyses

# Web scraping et IA générative : ce que change le projet de lignes directrices du CEPD

**Rédaction**

29 octobre 2026 · FR

**Résumé**

Le projet de Lignes directrices 03/2026 du CEPD place le web scraping destiné à l'IA générative au cœur des exigences du RGPD. Cette analyse examine l'accessibilité publique des données, la limitation des finalités, l'intérêt légitime, la minimisation, l'exactitude, la transparence, les données sensibles et l'importance croissante de la provenance des données dans la gouvernance de l'IA.

---

L'appétit considérable de l'intelligence artificielle générative pour les données a fait du web scraping une question juridique à part entière. Pages publiques, profils professionnels, réseaux sociaux, forums ou banques d'images offrent aux développeurs une matière accessible à une échelle sans précédent. Cette facilité technique ne dit pourtant rien, à elle seule, de la licéité de la collecte.

Le 8 juillet 2026, le Comité européen de la protection des données (CEPD) a adopté, en vue d'une consultation publique, ses Lignes directrices 03/2026 sur le web scraping dans le contexte de l'IA générative. La consultation reste ouverte jusqu'au 30 octobre 2026. Le texte n'est donc pas définitif. Il permet néanmoins de discerner assez nettement la manière dont les autorités européennes de contrôle entendent appliquer le RGPD aux collectes massives de données accessibles en ligne.

Le CEPD ne cherche pas à interdire le scraping. Son approche est plus exigeante : il demande au responsable du traitement de pouvoir justifier toute la chaîne, depuis la collecte des données jusqu'à leur utilisation dans le développement du modèle.

## Une donnée accessible au public n'est pas libre de droits

Le premier constat paraît simple, mais il emporte des conséquences importantes.

Dès lors que le scraping entraîne la collecte, l'enregistrement, l'organisation ou la consultation de données à caractère personnel, il constitue un traitement soumis au RGPD. La formule selon laquelle « ces informations étaient déjà publiques » ne suffit

donc pas à établir la licéité de leur réutilisation.

Cette distinction prend une importance particulière pour l'IA.

Une personne peut rendre son parcours professionnel accessible afin d'être identifiée dans son domaine, participer à un forum pour répondre à une discussion précise ou publier une photographie dans un contexte social déterminé. Rien ne permet d'en déduire automatiquement qu'elle pouvait raisonnablement anticiper l'aspiration massive de ces contenus en vue de constituer un jeu de données d'entraînement.

L'accessibilité publique est un élément de l'analyse. Elle ne dispense pas de l'analyse.

## Le scraping oblige à préciser ce que l'on cherche réellement à faire

Le principe de limitation des finalités occupe une place centrale dans l'approche du CEPD.

La collecte à grande échelle favorise facilement une confusion entre deux idées : pouvoir techniquement obtenir une donnée et avoir une raison juridiquement défendable de la collecter.

Une finalité aussi générale que « développer une intelligence artificielle » ou « améliorer un modèle » explique difficilement, à elle seule, pourquoi telle catégorie de données personnelles serait nécessaire.

Ce point n'a rien de théorique. La définition de la finalité conditionne le choix des sources, la quantité de données collectées, les catégories susceptibles d'être écartées, leur durée de conservation et, plus largement, l'appréciation de la proportionnalité du traitement.

Le web scraping réunit dans un même ensemble des informations issues d'univers très différents. Un profil professionnel, un message publié sur un forum et une photographie personnelle peuvent devenir de simples lignes dans un corpus d'entraînement. Le droit de la protection des données oblige au contraire à retrouver le contexte que l'agrégation technique tend à effacer.

La difficulté ne se résume donc pas à une opposition abstraite entre innovation et vie privée. Elle réside dans la confrontation de deux logiques : celle de l'agrégation massive et celle d'un droit qui exige de définir précisément la raison du traitement.

## L'intérêt légitime reste possible, mais il ne constitue pas un passe-droit

Pour les développeurs d'IA, la question de l'intérêt légitime au sens de l'article 6, paragraphe 1, point f), du RGPD est probablement l'un des aspects les plus importants du projet.

Le CEPD n'exclut pas cette base juridique pour l'entraînement d'un modèle à partir de données issues du web scraping. Une telle exclusion aurait constitué une position beaucoup plus radicale.

Il s'inscrit plutôt dans la continuité de son Avis 28/2024 relatif aux modèles d'IA.

L’intérêt invoqué doit d’abord être licite, clairement défini et réel. Il faut ensuite montrer que le traitement est nécessaire à sa poursuite. Enfin vient la mise en balance avec les intérêts, les droits et les libertés des personnes concernées.

C’est probablement cette troisième étape qui soulèvera le plus de difficultés en pratique.

Le caractère public d’une information compte, mais son contexte de publication compte également. Il peut être nécessaire d’examiner la nature du site source, le caractère plus ou moins sensible de l’information, l’existence d’une relation entre la personne et le responsable du traitement, les paramètres de confidentialité proposés par le service, les usages ultérieurs possibles du modèle et les attentes raisonnables de la personne.

L’intérêt légitime demeure donc une voie possible. Il ne transforme pas pour autant l’Internet ouvert en réserve générale de données d’entraînement.

Deux opérations de scraping poursuivant des objectifs techniques proches peuvent ainsi aboutir à des appréciations juridiques différentes en fonction des sources utilisées, des données collectées et des garanties mises en place.

## La minimisation des données face à la logique du « toujours plus »

Le développement des modèles génératifs s’est largement construit autour d’une intuition d’ingénierie : davantage de données permettraient souvent d’obtenir de meilleures performances.

Le RGPD part d’une autre exigence. Les données personnelles doivent être adéquates, pertinentes et limitées à ce qui est nécessaire au regard de la finalité poursuivie.

Ces deux logiques ne rendent pas nécessairement l’entraînement de modèles incompatible avec le droit européen. Elles obligent cependant les développeurs à effectuer des choix.

Le CEPD attend notamment que soit examinée la possibilité d’exclure certaines catégories de données ou certaines sources. Son Avis 28/2024 évoquait déjà des mesures consistant à écarter des sites particulièrement sensibles ou à tenir compte des mécanismes par lesquels des éditeurs s’opposent au crawling et au scraping automatisés.

La conformité se déplace ainsi vers l’amont.

Il devient difficile de soutenir une approche consistant à collecter d’abord le corpus le plus vaste possible, puis à laisser aux équipes juridiques le soin d’en réduire ultérieurement les risques. Le choix des sources et les mécanismes de filtrage font eux-mêmes partie de la conception juridique du système.

La gouvernance des données commence avant l’entraînement.

## L’exactitude ne concerne pas seulement les réponses du modèle

Les débats autour de l’IA générative mettent souvent l’accent sur les hallucinations et les informations erronées produites par le système. Le CEPD attire l’attention sur un moment antérieur : celui de la constitution du corpus.

Les informations disponibles en ligne peuvent déjà être anciennes, incomplètes, dupliquées ou privées de leur contexte.

Introduites à grande échelle dans un ensemble d'entraînement, ces erreurs deviennent difficiles à identifier et plus difficiles encore à corriger une fois qu'elles ont participé au développement du modèle.

Le CEPD recommande dès lors de privilégier des sources fiables, d'enregistrer le moment auquel les données ont été collectées et de procéder à leur validation avant utilisation dans l'entraînement.

La provenance et la qualité du jeu de données cessent ainsi d'être de simples préoccupations techniques. Elles rejoignent directement le principe d'exactitude du RGPD.

Autrement dit, les fondations informationnelles d'un modèle peuvent elles-mêmes devenir un objet de responsabilité juridique.

## La transparence ne disparaît pas avec la taille du corpus

Le web scraping se déroule, la plupart du temps, sans que les personnes concernées en aient conscience. À l'échelle des corpus utilisés pour l'IA générative, les informer individuellement peut également devenir extrêmement complexe.

Le CEPD ne méconnaît pas cette réalité.

Le RGPD prévoit lui-même certaines situations dans lesquelles la fourniture individuelle d'une information peut s'avérer impossible ou exiger des efforts disproportionnés. Mais cette exception ne peut être transformée en dispense générale de transparence pour les développeurs d'IA.

Le responsable doit toujours déterminer quelles obligations lui sont applicables et être en mesure de justifier, le cas échéant, le recours à une exception.

Lorsque l'information individuelle n'est pas exigée, la qualité des autres mécanismes de transparence prend davantage d'importance : notices publiques facilement accessibles, explication des pratiques de scraping, définition intelligible des finalités et dispositifs permettant aux personnes de comprendre ce qui peut être fait de leurs données.

La masse du traitement peut modifier les modalités de la transparence. Elle n'en efface pas le principe.

## Les données sensibles révèlent les limites d'une collecte indifférenciée

La difficulté devient plus nette encore lorsque les corpus aspirés contiennent des catégories particulières de données.

Opinions politiques, convictions religieuses, informations de santé, origine raciale ou ethnique, orientation sexuelle : toutes ces informations peuvent apparaître dans des contenus accessibles au public. Elles peuvent être recherchées volontairement, collectées accidentellement ou simplement noyées dans des volumes de données tels qu'il devient difficile de les identifier avant aspiration.

L’article 9 du RGPD n’en cesse pas pour autant de s’appliquer.

Le CEPD rappelle que le traitement exige alors à la fois une base juridique au titre de l’article 6 et l’une des conditions prévues par l’article 9, paragraphe 2. La présence d’une donnée sensible sur Internet ne crée aucune exemption générale.

Le Comité évoque à cet égard l’arrêt *GC et autres* (C-136/17) de la Cour de justice. Son raisonnement peut présenter un intérêt lorsque la collecte de données sensibles est incidente ou résiduelle. Il doit cependant être manié avec prudence : il ne supprime ni l’analyse au cas par cas ni l’obligation de mettre en œuvre des mesures techniques et organisationnelles adaptées.

Pour les grands corpus d’entraînement, ces garanties peuvent devoir intervenir dès la collecte ou immédiatement après celle-ci, et non uniquement au stade de l’exploitation du modèle.

## La question de fond devient celle de la provenance

Les Lignes directrices portent sur le scraping, mais leur portée dépasse largement le moment de la collecte.

C’est ici que leur articulation avec l’Avis 28/2024 du CEPD prend toute son importance. Lorsqu’un modèle a été développé au moyen d’un traitement illicite de données personnelles, cette histoire peut influencer l’appréciation juridique de traitements ultérieurs mettant en œuvre le modèle. Tout dépend des circonstances, notamment de la question de savoir si le modèle peut véritablement être considéré comme anonyme.

La provenance ne constitue donc plus simplement une fiche technique retraçant l’origine des données.

Un fournisseur d’IA peut être amené à expliquer quelles sources ont été utilisées, pourquoi elles ont été retenues, quelle base juridique fondait leur collecte, quelles exclusions ont été appliquées et quelles garanties ont protégé les personnes concernées.

Le modèle peut être mis sur le marché plusieurs mois ou plusieurs années après le scraping initial. Cela ne signifie pas que la phase de collecte a perdu toute pertinence juridique.

Documenter la chaîne de données devient ainsi un élément de gestion du risque réglementaire.

## D’une logique d’autorisation à une logique de responsabilité

Réduire les Lignes directrices 03/2026 à la question « le web scraping est-il légal ? » serait probablement passer à côté de leur apport principal.

Il n’existe pas de réponse unique.

Le texte dessine plutôt les contours d’un scraping dont la licéité doit pouvoir être démontrée. Il ne suffit plus de constater qu’une information était accessible. Il faut être en mesure d’expliquer pourquoi elle a été collectée dans ce contexte précis, pourquoi cette collecte était nécessaire et quelles mesures ont été prises pour en limiter les effets sur les personnes concernées.

Le débat se déplace ainsi.

L'Internet rend les données techniquement accessibles. Le RGPD pose d'autres questions : pour quelle finalité sont-elles reprises ? pouvait-on raisonnablement anticiper cette réutilisation ? quels risques en découlent ? quelles garanties ont été prévues ?

Pour les développeurs d'IA, ces interrogations entrent désormais dans la conception même des produits.

## Conclusion

Le projet de Lignes directrices 03/2026 ne ferme pas la porte au web scraping pour l'IA générative. Il rappelle en revanche que l'échelle de la collecte et le caractère public des données ne peuvent tenir lieu de justification juridique.

Le développeur qui souhaite s'appuyer sur des données personnelles accessibles en ligne devra de plus en plus pouvoir rendre compte de leur provenance, de la finalité de la collecte, de la base juridique retenue, des attentes raisonnables des personnes et des garanties intégrées au processus d'acquisition.

Le texte définitif pourra encore évoluer à l'issue de la consultation. Mais une orientation se dégage déjà nettement.

Le droit européen de la protection des données déplace progressivement son regard vers l'amont de la chaîne de l'IA : non plus seulement vers les résultats produits par le modèle, mais vers les choix qui ont permis de constituer ses données d'entraînement.

Dans ce contexte, la question posée à un fournisseur d'IA ne sera bientôt plus seulement de savoir s'il était techniquement possible d'aspirer les données. Il faudra pouvoir expliquer pourquoi il était juridiquement défendable de le faire.

### Sources

Comité européen de la protection des données, *Lignes directrices 03/2026 sur le web scraping dans le contexte de l'intelligence artificielle générative*, version adoptée pour consultation publique le 8 juillet 2026.

Comité européen de la protection des données, *Avis 28/2024 relatif à certains aspects de la protection des données liés au traitement de données à caractère personnel dans le contexte des modèles d'IA*, 18 décembre 2024.

Cour de justice de l'Union européenne, *GC et autres c. CNIL*, C-136/17, arrêt du 24 septembre 2019, ECLI:EU:C:2019:773.

Règlement (UE) 2016/679 (RGPD), notamment ses articles 5, 6, 9, 13 et 14.

### Comment citer cet article

Rédaction, « Web scraping et IA générative : ce que change le projet de lignes directrices du CEPD », JurisPONt, 29 octobre 2026,
<https://jurispont.com/fr/articles/web-scraping-ia-generative-lignes-directrices-cepd-2026>