

● **IA générative : le CEPD adopte des lignes directrices sur le web scraping pour l'entraînement des modèles**

Le 7 juillet 2026, le Comité européen de la protection des données (« CEPD ») a adopté des lignes directrices 03/2026 relatives au moissonnage de données sur internet (« web scraping ») dans le contexte de l'intelligence artificielle générative (« IA générative »). Le texte précise les conditions dans lesquelles la collecte automatisée de données accessibles sur internet à des fins d'entraînement de modèles d'IA générative peut être réalisée, conformément au règlement général (UE) 2016/679 sur la protection des données (« RGPD »).

Le CEPD définit le web scraping comme une technique reposant sur des outils automatisés permettant d'extraire et de stocker des informations à partir de services web accessibles au public, tels que des registres publics, des portails open data, des médias en ligne, des réseaux sociaux ou des forums de discussion.

Les lignes directrices précisent les règles à respecter lors de la collecte de données sur internet pour entraîner des modèles d'IA générative, en particulier s'agissant de la base légale, de l'information des personnes concernées, de la minimisation des données collectées et du traitement des données sensibles. Elles distinguent le web scraping ciblé, réalisé selon des critères de collecte prédéfinis, du web scraping non ciblé, dans lequel un outil explore et suit automatiquement les liens découverts, avec le risque de collecter un volume important de données sans que le responsable du traitement maîtrise précisément la nature des données recueillies.

I. Sur le choix de la base légale applicable au web scraping

Le CEPD relève que l'intérêt légitime prévu à l'article 6, paragraphe 1, point f), du RGPD est fréquemment invoqué par les acteurs pour les opérations de web scraping destinées au développement de modèles d'IA générative. Il souligne en revanche qu'il sera généralement difficile de fonder la collecte sur le consentement, compte tenu du caractère indirect et massif de la collecte. La seule mise à disposition publique de données sur internet ne signifie d'ailleurs pas que la personne concernée a consenti à leur collecte à des fins d'entraînement d'un modèle d'IA.

Le recours à l'intérêt légitime suppose le respect de trois conditions cumulatives : (i) l'existence d'un intérêt légitime poursuivi par le responsable du traitement ou un tiers, (ii) la nécessité du traitement pour atteindre cet intérêt et (iii) la mise en balance de cet intérêt avec les droits et libertés des personnes concernées.

Sur la nécessité du traitement, le CEPD invite notamment les responsables du traitement à déterminer si la finalité poursuivie pourrait être atteinte par un moyen moins intrusif. Une collecte plus ciblée, le recours à des données pseudonymisées ou l'utilisation de données de synthèse peuvent constituer des alternatives à une collecte extensive de données personnelles.

Dans le cadre de la mise en balance, doivent notamment être pris en compte la nature des données, l'impact du traitement sur les personnes concernées et leurs attentes raisonnables. Le CEPD souligne que le fait qu'une information soit accessible en ligne ne signifie pas que les personnes doivent nécessairement s'attendre à ce qu'elle soit collectée et réutilisée pour l'entraînement d'un modèle d'IA, dans toutes les situations et pour toutes les finalités.

II. Sur l'encadrement des données sensibles moissonnées en ligne

Le traitement de données sensibles est, en principe, interdit par l'article 9 du RGPD. Lorsque le web scraping porte intentionnellement sur de telles données, le traitement doit, en plus de reposer sur une base légale au titre de l'article 6 du RGPD, relever de l'une des dérogations prévues à l'article 9, paragraphe 2.

Le responsable du traitement doit prendre des mesures pour éviter de collecter ces données. Toutefois, il est souvent difficile de savoir avec une certitude raisonnable, avant le web scraping, si celui-ci en collectera. Une collecte résiduelle de telles données est donc probable.

Pour tenir compte de cette difficulté, le CEPD s'appuie notamment sur l'arrêt GC e.a. du 24 septembre 2019 (C-136/17), dans lequel la CJUE a précisé les obligations d'un moteur de recherche lorsqu'il référence des pages contenant des catégories particulières de données. La CJUE juge que l'interdiction de traiter des données sensibles s'applique à l'exploitant d'un moteur de recherche dans la mesure de ses responsabilités, de ses compétences et de ses possibilités, compte tenu du rôle spécifique qu'il joue dans le référencement des pages contenant ces données.

Le CEPD précise les mesures que le responsable du traitement doit mettre en œuvre lorsqu'il existe un risque de collecter ou de restituer des données sensibles. Avant la collecte, le responsable du traitement doit définir des critères précis, appliquer des filtres destinés à éviter la collecte de données sensibles et exclure les catégories de sites qui, a priori, en contiennent. Les données sensibles qui auraient néanmoins été collectées doivent ensuite être supprimées dès leur identification. Pendant le développement du modèle, le responsable du traitement doit mettre en place des mesures pour éviter que le modèle ne restitue ou ne révèle des données sensibles présentes dans les données d'entraînement. Une fois le modèle développé, le responsable du traitement doit surveiller les réponses générées par le système d'IA et intervenir si celles-ci révèlent des données sensibles.

Lien utile :

- [Lignes directrices 03/2026 sur le web scraping dans le contexte de l'IA générative](#)