Scraping de données et régularité

Thomas Saint-Aubin, Charles Leconte

scraping de données et légalité

Si les données constituent l'un des principaux actifs immatériels des entreprises, la légalité du processus de collecte automatisée est un préalable indispensable pour pouvoir valoriser ultérieurement ce patrimoine informationnel.

Avant de pouvoir exploiter les données, il faut les collecter. Il existe différents moyens de collecter des data : soit manuellement, ce qui peut nécessiter un temps considérable lorsqu'on cherche à disposer d'un volume important, soit au moyen de méthodes automatiques, via des logiciels permettant d'obtenir une quantité importante d'informations dans un laps de temps record (action de *« scraper »*).
De plus en plus de sociétés ont recours au *« web scraping »* pour récupérer le contenu des sites afin d'enrichir leur propre base ou de générer de nouveaux business. La question de la légalité de cette activité est souvent traitée tardivement, notamment par les investisseurs dans le cadre de la *« due diligence »* (diligence raisonnable) ou encore au moment de la démarche de mise en conformité RGPD.

scraping et non crawling

« Scraping » est un terme anglais signifiant littéralement *« grattage »*, qu'il ne faut pas confondre avec le *« web crawling »*, pratique d'indexation des données sur les moteurs. Appliqué au web, le terme renvoie à une technique d'extraction automatisée de contenu structuré. Concrètement, la récupération de données est effectuée par un programme, un script qui va parcourir un site web et extraire les données et les stocker dans le but de les réutiliser sur son propre site.
Mais comment le droit qualifie-t-il et encadre-t-il la collecte des données, particulièrement lorsqu'elle est opérée par des robots ?

1. Scraping vs mise à disposition d'API

Le scraping est différent de l'usage d'une interface de programmation applicative (API) permettant au site source de contrôler le transfert des données aux tiers réutilisateurs en fournissant un accès gratuit ou payant.
La méthode la plus courante pour réaliser un scraping légal de données est de recenser et de récupérer des données publiques distribuées sous une licence libre et ouverte. En France il s'agira nécessairement de l'une des licences énumérées dans le décret n° 2017-638 du 27 avril 2017 relatif aux licences de réutilisation de l'open data.
Mais au moment des débats sur la création d'un service public de la donnée, consacrée par le décret n° 2017-331 du 14 mars 2017, une partie des praticiens se prononçait pour la création d'un service public de mise à disposition des API de données publiques de référence. Malheureusement, cette proposition n'a pas été imposée comme un corollaire obligatoire pour ces données essentielles. En l'état, c'est donc à chacun des acteurs concernés de financer et de mettre en place des pratiques de scraping des données publiques plutôt que de faire reposer cet investissement initial et mutualiser sur les principaux producteurs de données publiques.

2. métadonnées juridiques associées à un jeu de données

Le scraping connaît un regain d'intérêt depuis 2010 avec l'apparition des activités de *« growth hacking »*. La constitution de base de données de prospects et le scraping des réseaux sociaux est une pratique courante chez les growth hackers. Comment encadrer juridiquement la récupération et la réutilisation des données privées ?
Les derniers travaux collaboratifs portés par l'écosystème Privacy Tech, notamment initiés dans le cadre de Design Your Privacy, cherchent à créer un référentiel des CGR (pour passer des *« conditions générales d'utilisation »* aux *« conditions générales de réutilisation »*) afin de redonner concrètement le contrôle aux personnes concernées par les traitements. Ces travaux s'intègrent dans une démarche plus générale de création de standards juridico-techniques pour

concrétiser la reconnaissance d'un droit à la portabilité des données.
Ils seront présentés à l'Assemblée nationale le 10 avril 2019 : un consortium européen est en cours de constitution.

« associer directement à des données une synthèse des droits et permissions »

projet « license your data »

Le projet porté par l'Institut national de recherche dédié aux sciences du numérique (Inria), *« license your data »*, envisage plus globalement d'associer des métadonnées juridiques à chaque jeu de données disponibles sur le web des données.
Pour éduquer le robot au scraping légal des données, c'est probablement à moyen terme la bonne pratique à encourager : ce sera à la personne concernée ou au producteur de proposer une version codée du droit applicable et de la licence associée.
Cette pratique, en plein développement dans le cadre des créations de *« data lakes »* dans les grandes entreprises, permet d'associer directement à des données une synthèse des droits et permissions.
Si les dernières innovations en matière de legaltech et Privacy Tech permettent d'envisager une concrétisation du *« Law is code »* dans le droit des données et les stratégies de valorisation des données, faut-il modifier le cadre juridique applicable ?

3. état du droit applicable au scraping

À l'ère du big data et des pratiques généralisées du scraping des données, Me Nicolas Courtier (2) remet en cause la protection sui generis des producteurs des bases de données de la loi de 1998 : *« Le droit des producteurs des bases de données repose sur une approche statique des traitements de données : on se concentre sur la création de la base et non sur son utilisation »*.
En droit positif, plusieurs textes (en propriété intellectuelle, secret des affaires, protection des données personnelles, action en concurrence déloyale, etc.) permettent de sanctionner le scraping illégal.
En propriété intellectuelle, le droit sui generis (art. L342-1 du CPI) permet au producteur de la base de données d'interdire, entre autres, l'extraction, par transfert, ou la réutilisation de la totalité ou d'une partie du contenu de sa base de données.

délit de vol de donnée

En droit pénal, le législateur a fait de l'extraction de données un délit spécifique. La loi du 24 juillet 2015 a modifié l'article 323-3 du Code pénal qui réprime désormais le fait *« d'extraire, de détenir, de reproduire, de transmettre »* frauduleusement les données d'un système de traitement automatisé de données (STAD). Le vol de donnée est donc bien un délit distinct du vol d'une chose matérielle (art. 311-1 C. Pénal).
Au-delà de l'idée d'une réforme du droit positif applicable portée par Me Courtier, nous pensons que c'est d'abord une application combinée du droit et de la technique qui permettra d'encadrer juridiquement le scraping de la donnée.
La technologie blockchain, via des licences d'API ou associées à des jeux de données transformées en smart contracts, permet déjà de *« tokeniser »* les licences, de tracer les réutilisations des données et d'en répartir automatiquement le fructus.
Si le droit permet aux entreprises de ne pas être tout à fait démunies en matière de protection de leur patrimoine informationnel et que la legaltech progresse rapidement, elles peuvent également prévenir la récupération de leurs données en amont en protégeant leur site (création de compte utilisateur, bannissement d'IP, captchas, etc.)

4. retour d'expérience sur une pratique de scraping légal

La propriété intellectuelle connaît un gisement considérable de données librement réutilisables. Dans cette matière, plusieurs institutions distribuent des jeux de données, allant de la jurisprudence PI aux titres de droit de PI.
Notre enjeu était de mettre en place un robot juriste scrapeur en capacité de collecter légalement ces données pour pouvoir ultérieurement les exploiter. Nous avons fonctionné en quatre étapes :
❶ recensement des jeux de données disponibles à l'international en matière de données de jurisprudence PI ;
❷ recensement des licences associées et interrogation des institutions productrices (le cas échéant) pour connaître le cadre applicable à la réutilisation ;
❸ modélisation des métadonnées juridiques associées aux jeux de données récupérées ;
❹ création du robot scrapeur pour collecter les données, modéliser les métadonnées juridiques associées, enrichir sémantiquement les données récupérées et les intégrer à la base.
Ce projet est en ligne sur Caseip.com (3). Il intègre aujourd'hui les données de jurisprudence française et des institutions européennes. Dans les prochains mois, la plateforme proposera l'accès centralisé aux données de jurisprudence PI allemandes et américaines. Pour collecter cette donnée, le robot scrapeur est en plein apprentissage ! ■

Thomas Saint-Aubin
[Juriste, chercheur en entrepreneur, administrateur de l'Adij, fondateur de Seraphin.legal, le studio Legal Tech]
→ www.seraphin.legal

Charles Leconte
[Juriste et cofondateur de CaseIP]
→ www.caseip.com

(1) Inria : Institut national de recherche dédié aux sciences du numérique.

(2) Avocat au barreau de Marseille, spécialiste en droit de la propriété intellectuelle et droit des nouvelles technologies, de l'informatique et de la communication.

(3) → www.caseip.com