La collecte automatisée de données : Crawling & Scraping

Actualité législative publié le **06/04/2018**, vu **10131 fois**, Auteur : Murielle Cahen

Aujourd'hui, il est indéniable que les nouvelles technologies prennent une place de plus en plus importante dans notre quotidien. Au regard de la production massive de données personnelles qui en découle, la question se pose de savoir comment encadrer leur collecte, notamment lorsqu'elle est automatisée, comme c'est le cas des pratiques de « crawling » et de « scraping ».

Ces logiciels permettent en effet, dans un laps de temps très court, d'obtenir une quantité importante d'informations utiles pour une entreprise ou un particulier, à partir d'une liste de sites constituant le « champ d'action » du robot.

Néanmoins, ces pratiques demeurent encadrées. Elles doivent répondre à certains principes, et notamment à ceux liés à la protection des données collectées automatiquement.

Dès lors la propriété, la nécessité d'une autorisation préalable pour la collecte, ou encore les questions liées à la réutilisation de ces données sont des enjeux de taille qui dictent les limites de la légalité de ces outils de collecte automatisée.

Pour en saisir toute l'importance il convient donc de comprendre, dans un premier temps, les différents usages qui peuvent être faits de ces outils (I), pour ensuite envisager le cadre protecteur des données collectées automatiquement (II).

1. ## Les différents usages des crawlers et scrapers

La récolte des données à des fins d'information (A), tout comme l'indexation et la réutilisation de celles-ci (B), sont les objectifs visés par l'usage de ces outils numériques.

1. ### La récolte des données

Le *crawling* est une pratique qui consiste à « *collecter [automatiquement] le contenu d'une page pour ensuite la traiter, la classer et fournir des informations* » au propriétaire du logiciel.

Le logiciel de *scraping*, lui, va « *extraire du contenu d'un site Web dans le but de le transformer pour permettre son utilisation dans un autre contexte* ».

Néanmoins, la récolte de ces données ne va pas fonctionner sur le même principe, que l'on soit dans le cas des *crawlers* ou dans celui des scrapers.

En effet, les *crawlers* vont fonctionner sur un principe de redirection : le logiciel va dans un premier temps se rendre sur des pages prédéfinies pour en récupérer l'intégralité du contenu. Par la suite, il va extraire l'ensemble des liens URLs présents sur les pages analysées, et suivre ces liens pour également analyser le contenu des pages référencées sous ces liens.

Le scraper, lui, va plutôt se baser sur un « patron » configuré au préalable, qui prend en compte la structure HTML de la base de donnée analysée, afin de pouvoir extraire de manière pertinente les données et leur mise à disposition sur les pages consultées.

2. ### L'indexation et la réutilisation des données

Des questions peuvent se poser au regard de l'exploitation des données récoltées par ces outils.

L'objectif principal demeure celui de tirer des informations pratiques et concrètes de ces données : une fois récoltées, puis triées et structurées en fonction de leur pertinence et de ce que recherche l'auteur, elles permettront d'avoir une vision précise du contenu et des pratiques, pour l'usager, des pages analysées.

Mais, comme on l'a vu, ces données peuvent également être réexploitées dans un but bien précis : c'est l'exemple de la plateforme américaine *Common Crawl*, ayant pour objectif d'archiver le plus de pages Web possible, et de rendre disponible leur accès via le site de la fondation. On estime qu'aujourd'hui, la plateforme centralise environ 15 % du web mondial, grâce à l'usage de *crawlers*.

De plus, certains pourraient être tentés de réutiliser les données collectées, afin par exemple d'augmenter le trafic de leur propre site internet.

Ces pratiques posent plusieurs questions, et notamment au regard du droit de la propriété intellectuelle et de la protection accordée à ces données et bases de données.

2. # Les atteintes à la protection de ces données

La propriété intellectuelle et le droit d'auteur offrent un cadre légal protection aux données récoltées automatiquement (A). Ceci étant, le propriétaire de ces données pourra également chercher à se prémunir lui-même d'une telle collecte (B).

1. ### Le cadre imposé par le droit de la propriété intellectuelle et le droit d'auteur

Il faut savoir que ces pratiques sont encadrées par le droit, et notamment par la propriété intellectuelle, pour éviter tout type d'abus et notamment la contrefaçon.

Dans le cadre d'une indexation des données, en réalité, la contrefaçon ne sera généralement

pas admise si les sources sont référencées.

C'est notamment ce qu'a pu retenir le Tribunal de grande instance de Paris, dans son arrêt « *Adenclassified* » du 1er février 2011 ayant débouté de sa demande une société dont les données ont été indexées, les faits ne constituant pas une violation du « *droit sui generis du producteur de bases de données* ».

À contrario, l'extraction de données par le biais de ces outils numériques dans la poursuite d'un objectif de réutilisation « *de la totalité ou d'une partie qualitativement ou quantitativement substantielle du contenu d'une base de données* » est constitutive d'un acte de contrefaçon, comme le prévoient expressément les articles 342-1 et 342-2 du Code de la propriété intellectuelle.

Au demeurant, il n'existe pas de règles précises concernant l'établissement du caractère substantiel du contenu. Ainsi, la reconnaissance d'un tel critère se fera au cas par cas par le juge du litige en question.

2. Les moyens de lutte contre ces outils

Il est souvent recommandé aux utilisateurs de *crawlers* et *scrapers* d'agir avec mesure et parcimonie : par exemple, ceux-ci ne devront pas surcharger les serveurs des sites visités par un nombre de requêtes trop important, au risque de causer un déni de service qui pourra facilement s'apparenter à un acte de concurrence déloyale.

En outre, certains propriétaires de sites peuvent se prémunir face à ces outils, refusant de voir leurs données récoltées « *pillées* ».

La Cour d'appel de Paris, dans son arrêt « *SAIF c/Google* » du 26 janvier 2011, soutenait effectivement que « *chaque webmaster peut, via son fichier robot.txt, contrôler la manière dont les données de son site sont visitées par les crawlers, notamment en interdisant l'accès à certaines d'entre elles* ».

La légalité, tout comme la légitimité, du *crawling* et du *scraping* restent donc encore aujourd'hui discutables.