La protection des données personnelles à l'ère du big data

Le big data semble se nourrir insatiablement de toutes nos données, personnelles ou non. Un « monstre » qui représente un danger de taille : il est non seulement en mesure de porter atteinte à notre vie privée, mais, au-delà, capable aussi d'agir sur notre libre arbitre.

Maryse ARTIGUELONG, coresponsable du groupe de travail LDH « Libertés et Tic »

« *Données structurées ou non, dont le très grand volume requiert des outils d'analyse adaptés* ». C'est la définition que donne la Commission générale de terminologie et de néologie, qui préconise depuis août 2014 l'utilisation du terme « mégadonnées » (ou encore « données massives ») en lieu et place de « big data ».

Le big data désigne en effet des volumes importants de données très diverses, traitées et analysées pour extraire des informations qui seront utilisées dans de nombreux domaines. Ces mégadonnées sont qualifiées de « carburant » ou d'« or noir » car leur valeur alimente l'économie numérique. Leur exploitation permet aussi bien d'identifier les causes endogènes et exogènes des maladies (en particulier du cancer) que de fluidifier le trafic routier, de préciser le ciblage publicitaire, de réduire la consommation énergétique.

La disponibilité de ces données est facilitée par l'évolution des techniques de transport et de stockage (*cloud*, fibre, etc.), qui facilitent la croissance exponentielle de la rétention de « toutes » les données.

Les « 3 V » : Volume, Variété, Vélocité

Le big data est souvent caractérisé par les « 3 V » de Volume, Variété, Vélocité.

Le « Volume », soit la quantité de données à disposition, est colossale. Elle se composerait de plus de mille deux cent milliards de milliards d'octets, dont 90 % ont été produits dans les deux dernières années, et ce chiffre devrait doubler tous les deux ans. Le volume joue un rôle capital, en effet la justesse des informations dégagées dépend de la quantité de données traitées. On ne cherche plus à travailler sur un échantillon mais sur tous les cas réels d'un phénomène donné. On comprend alors le puissant mouvement de recueil des données, qui s'accommode mal de la nécessité d'attendre des consentements individuels.

La « Variété » du big data caractérise les formats hétérogènes des données provenant de sources très diverses : téléphones mobiles, téléviseurs connectés, tablettes, PC fixes, PC portables, et, de plus en plus, objets connectés. Elles sont délivrées en toute légalité par les utilisateurs de services, conservées éventuellement au-delà de la finalité fixée, « récupérées » grâce aux traces de navigation sur le Net, aux GPS, réseaux sociaux, objets connectés, aux métadonnées [1] mais aussi aux « traceurs » que sont les cartes bancaires, cartes de fidélité, cartes d'accès aux transports… Les données publiques peuvent être des données administratives comme des registres de naissance, des listes électorales mais aussi des données scientifiques. Les données privées sont recueillies sous les formats textes, images, sons, traces de navigation sur Internet.

> ***« La quantité de données à disposition est colossale : plus de mille deux cent milliards de milliards d'octets, dont 90 % ont été produits dans les deux dernières années. Un chiffre qui devrait doubler tous les deux ans. »***

La « Vélocité » caractérise à la fois la rapidité de production et de traitement des mégadonnées. C'est le « *data mining* » [2], qui permet de traiter rapidement les masses de données grâce aux algorithmes. Ces outils d'analyse, de plus en plus puissants, sont capables de s'autocorriger (« *machine learning* »), gèrent

(1) Ce sont les données (ou informations) concernant une donnée. Pour un message il s'agira de la date, du lieu d'envoi, du destinataire, de l'expéditeur et objet.

(2) Forage ou exploitation des données.

des informations qui n'ont plus besoin d'être structurées dans des bases de données, et permettent de détecter des relations entre des données hétérogènes provenant de différents contextes.
Les algorithmes permettent d'extraire des informations, de définir des profils types, non seulement de repérer des comportements suspects mais aussi de les prédire, tout comme sont anticipés des événements ou des tendances : vendre des produits avant de les fabriquer, connaître à quel moment l'internaute sera prêt à passer à l'acte d'achat… Mais ils permettent aussi d'accélérer le séquençage du génome humain ou d'établir une cartographie des différentes formes de maladies dégénératives du cerveau.

Le big data en pratique

Les avantages de l'exploitation du big data sont importants, comme le montre l'exemple des transports.
En 2012, une équipe de chercheurs de la société IBM a exploité une base de données mise à disposition par l'opérateur Orange. La société a recueilli et sauvegardé les données téléphoniques correspondant à cinq cent mille appels et SMS envoyés durant cinq mois, dans la ville d'Abidjan. Ces enregistrements, comprenant un identifiant anonymisé [3], l'heure de l'appel ou de l'envoi du message ainsi que l'identifiant de l'antenne-relais du début de connexion ont permis aux chercheurs d'établir la carte des déplacements et de formuler des recommandations pour optimiser la carte des transports en commun. L'objectif : diminuer le temps de trajet quotidien de millions de personnes.
Une étude similaire, utilisant les données anonymisées des cartes de transport, a été réalisée sur les trajets quotidiens à Londres. Elle permet d'anticiper la congestion des bus et des métros et d'informer les usagers par le biais de comptes Twitter ; bientôt des informations en temps réel pourront être fournies pour leur permettre d'adapter leurs trajets.
S'il s'agit dans les deux cas de données anonymisées, celles-ci sont utilisées à une autre fin que celle initialement prévue… Mais le plus préoccupant concerne le respect de la confidentialité des données personnelles. En 2006, le site « Netflix », dans le cadre d'un concours pour améliorer son système de recommandation, a publié les choix en ligne d'un demi-million d'utilisateurs identifiés par un simple numéro. Deux chercheurs ont pu ré-identifier plusieurs clients par simple recoupement avec les données publiées sur un autre site d'avis en ligne, « IMDb » (« Internet Movies Data base »), qui, lui, n'était pas anonyme ; et même, dans certains cas, déterminer leurs opinions politiques et leurs orientations sexuelles.
C'est aussi grâce au big data, en utilisant notamment les données personnelles de ses clients, qu'Amazon a mis au point un outil lui permettant de leur adresser des marchandises avant même qu'ils ne les aient commandées, ce qui suppose un profilage extraordinairement intrusif de leur intimité. De nouvelles pratiques de vente émergent, telles que la « tarification dynamique » (« *dynamic pricing* »), pratiquée par des compagnies aériennes, qui peuvent augmenter leurs tarifs en fonction de l'analyse du comportement du client, de son besoin plus ou moins urgent et des prix de la concurrence.
Au-delà des risques de profilage et d'atteinte à la vie privée, le big data, en anticipant sur les décisions des individus et en les « aidant » à consommer, à améliorer ou surveiller leur santé… est bien plus qu'un outil, il interfère sur leur libre arbitre et leur autodétermination. De plus les algorithmes, qui sont conçus par des individus, peuvent comporter des failles dues à leurs jugements de valeur. Il est donc nécessaire d'être prudent avec les résultats, qui pourraient en être biaisés.
Pour les défenseurs des droits de l'Homme attachés à la vie privée, le big data constitue un changement de paradigme car les différents principes de la protection des données personnelles ne sont plus respectés. Le droit à la protection des données personnelles est garanti par différents textes : en France, la loi Informatique et Libertés ; au niveau européen, principalement la directive 95/46/CE ainsi que la Convention 108 [4]. Objectif : établir un équilibre entre l'individu et la personne physique ou morale, l'autorité publique ou autre qui collecte et traite ses données.

> ***« Il est évident que tout nouveau traitement effectué à partir de données collectées pour une finalité explicite ne devrait être possible que sur la base du consentement de la personne concernée ou de la garantie que ses données anonymisées le resteront. »***

Les normes de protection remises en question

La norme de la protection des données implique le respect de plusieurs principes qui sont mis à mal par le big data. Ces principes édictent que le recueil des données doit correspondre à une finalité déterminée dès la collecte, et que la personne concernée doit avoir donné son consentement « *spécifique, libre, explicite et éclairé* ». Les données doivent être minimisées, adéquates, pertinentes et non excessives par rapport aux finalités pour lesquelles elles sont collectées et pour les éventuels traitements ultérieurs [5]. Elles ne sont conservées que pour une durée n'excédant pas celle nécessaire aux finalités pour lesquelles elles sont enregistrées.
Le big data implique a contrario que toutes les données possibles soient recueillies, conservées et réutilisées pour d'autres finalités, jamais effacées ou presque, sans que l'on ait demandé à la personne concernée son consentement pour cette nouvelle utilisation. Si les textes prévoient la réutilisation, c'est uniquement à des fins statistiques ou de recherche scientifique, ce qui

(3) Orange avait fait un travail important d'anonymisation et d'agrégation des données afin que l'on ne puisse en aucun cas remonter aux individus (avec destruction des données après utilisation). Des universitaires sollicités pour hacker les données n'avaient pas réussi à les désanonymiser.

(4) Convention pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel du Conseil de l'Europe, signée par quarante-huit, pays et qui a valeur de traité. http://conventions.coe.int/Treaty/fr/Treaties/Html/108.htm.

(5) Loi Informatique et Libertés, art. 6 - 2° : « *Elles sont collectées pour des finalités déterminées. […] Toutefois, un traitement ultérieur de données à des fins statistiques ou à des fins de recherche scientifique ou historique est considéré comme compatible avec les finalités initiales de la collecte des données, […] et s'il n'est pas utilisé pour prendre des décisions à l'égard des personnes concernées.* ».

© LICENCE CC

implique qu'elles soient anonymisées et que cette réutilisation ne serve pas à prendre des décisions à l'égard des personnes. Or de nombreux exemples montrent que des données anonymisées peuvent donner lieu à une ré-identification, que le big data peut servir à rejeter des candidats à l'embauche « grâce » à une analyse de leur vie numérique, à adapter des tarifs d'assurance au niveau de risque de certains assurés, patients...

L'urgence de sensibiliser et d'informer

Il est indéniable que le big data peut contribuer à de nombreux progrès, notamment par l'utilisation de données publiques. Les perspectives sont immenses et les pouvoirs publics y voient des possibilités d'économies ; les entreprises, des profits potentiels.

Néanmoins, il est inacceptable que le citoyen ne soit plus en mesure de faire valoir son droit à la protection. Lorsqu'il accepte que ses données soient collectées pour une finalité, il doit avoir la garantie qu'elles ne seront pas cédées ou vendues à une entreprise ou une autorité qui en fera un tout autre usage. C'est la base de la confiance, dont on nous redit qu'elle est le moteur de l'économie numérique (les données en étant le carburant...).

Or, à l'heure actuelle, ce citoyen peut rester dans l'ignorance de la cession pour d'autres usages de ses données collectées. Il risque ainsi d'être, un jour, victime d'une ré-identification, d'une discrimination liée au profil qui aura été établi grâce au big data. Il est évident que tout nouveau traitement effectué à partir de données collectées pour une finalité explicite ne devrait être possible que sur la base du consentement de la personne concernée ou de la garantie que ses données anonymisées le resteront.

Dans ce domaine, des progrès importants sont nécessaires. Il est du devoir de l'Etat d'encourager les recherches sur l'anonymisation irréversible et, lorsque celle-ci s'avère impossible, les données concernées (notamment en matière de santé) devraient être exclues du big data [6].

Parallèlement, il convient de lancer très rapidement une grande campagne de sensibilisation des citoyens aux enjeux de la constitution de ces réserves de mégadonnées et des utilisations qu'elles permettent. Il est par ailleurs nécessaire d'encourager les citoyens, par tous les moyens, à se protéger, par des formations dès l'école et par des informations facilement accessibles.

Au-delà des risques de profilage et d'atteinte à la vie privée, le big data, en anticipant sur les décisions des individus, est bien plus qu'un outil, il interfère sur leur autodétermination.

En effet, lorsque le consentement de l'utilisateur est lié à l'utilisation de ses données comme condition d'accès à une application ou un service, il est généralement conditionné par le confort du service immédiat. L'utilisateur ne fera pas la démarche de refuser l'enregistrement de ses données. Il ne fera pas plus l'effort d'effacer leurs traces, parce qu'il considère que cette activité n'est pas très importante (« *acheter un billet de train n'a rien de secret ! Et puis, je n'ai rien à cacher !* »). Il ne voit pas à quoi ses données, prises séparément, pourraient servir. Il a tendance à faire confiance à l'expert, qui a créé la règle par défaut, a priori pour son bien... Par ailleurs, la plupart des utilisateurs pensent qu'ils perdront des « avantages », s'ils ne consentent pas à délivrer leurs données. Beaucoup sont conscients des risques, mais ils se sont résignés à perdre le contrôle sur leurs données. Il est donc urgent de renverser cette tendance. ●

(6) Le projet de loi sur le numérique devrait y contribuer, en ajoutant aux missions de la Cnil le soutien au développement des technologies protectrices de la vie privée. A condition que les moyens nécessaires lui soient affectés... Le projet prévoit aussi d'obliger les plateformes Web à donner « *une information loyale, claire et transparente sur les conditions générales d'utilisation* » et à prévoir que le consommateur puisse disposer d'une fonctionnalité gratuite permettant la récupération licite de tous ses fichiers mis en ligne et de toutes les données associées au compte utilisateur.