Big Data et cybersécurité : quelles armes face à cette ruée vers ce nouvel or numérique ?

Publié le 5 octobre 2017 – Mis à jour le 26 janvier 2018

La donnée devient le « nouvel or » de nos sociétés modernes, une chance sans précédent pour élever significativement notre connaissance collective dans les domaines de la santé, de l'écologie ou encore de la cybersécurité à l'ère du *Big Data*. Cette richesse collective s'obtient par le croisement d'énormes banques de données (BD) qui peuvent révéler lors de ces corrélations des données à caractère personnel (DCP). Se pose alors la question de prévenir le risque de leur mésusage.

L'obtention indue de DCP constitue au regard de la loi française une violation des droits fondamentaux. Ainsi, l'article 4 du règlement européen a défini en avril 2016 la DCP : « toute information se rapportant à une personne physique identifiable celle-ci pouvant l'être directement ou indirectement, notamment par référence à un ou plusieurs identifiants (nom, numéro d'identification, localisation) ou éléments d'identification spécifiques à une identité physique, physiologique, génétique, psychique ». Le stockage d'une DCP a la structure suivante :

Nom	Prénom	Date de Naissance	Code Postal	Pathologie
Dupont	Raymond	12/07/1950	71100	Insuffisance rénale

Tab 1 – Illustration BD « A », 1 enregistrement non anonymisé

Dans cet exemple, l'obtention par un organisme bancaire de la DCP « nom » et « pathologie » pourrait compromettre l'accès à un prêt. Pour éviter cela, la loi encadre la manipulation et le stockage de données massives en recommandant

avant traitement la pseudonymisation des DCP. La pseudonymisation consiste à remplacer les attributs directement identifiants, comme le nom de la personne, par un pseudonyme tout en gardant l'information utile :

N°	Pseudonyme	Code Postal	Date de naissance	Pathologie
1	X	71100	12/12/1950	Insuffisance rénale
2	Y	69740	06/02/1965	pneumopathie
3	Z	69120	21/01/1965	pneumopathie
4	....	...	...	...

Tab 2 – BD « B » : 3 enregistrements pseudonymisés

Or le risque d'une ré-identification est particulièrement présent à l'ère du *Big Data* et de l'*Open Data*, ces deux phénomènes facilitent le croisement avec d'autres banques de données, moins maîtrisées.

Site Web des professionnels de santé

N°	Nom	S	Code Postal	Date naissance
1	Marron	H	71100	30/03/1950
2	Amery	H	60240	03/02/1975
3	Durant	H	68120	23/12/2006
4	Dupont	H	71100	12/12/1950
5	....		...	...

Tab 3 – Illustration de la BD « C »

N°	Nom	S	Code Postal	Date naissance
1	W	H	Saone et Loire	[65-70]
2	X	H	Oise	[40-45]
3	Y	H	Bas-Rhin	[10-15]
4	Z	H	Saone et Loire	[65-70]
5	....		...	...

Tab 4 – Illustration de la BD « C » anonymisée

Dans cet exemple, le patient « Dupont » est ré-identifié par le croisement des enregistrements {1} de la BD « B » et {4} de la BD « C » qui n'est qu'une simple banque de données de patients d'une application de rendez-vous médicaux en ligne. L'anonymisation est la meilleure technique pour garantir la vie privée tout en préservant l'utilité des DCP. Dans l'exemple de la BD « C » anonymisée, les DCP des personnes 1 et 4 correspondent. Ainsi, l'anonymisation diminue le risque de ré-identification mais ne l'exclut pas.

La cybersécurité une aide ou un risque pour l'anonymat ?

La cybersécurité est efficace dans l'analyse de traces informatiques laissées par les usagers, auteurs d'activités normales, malveillantes ou erronées. Lors d'incidents de sécurité, les opérateurs investiguent ces traces afin de leur attribuer une responsabilité civile ou pénale. Ces traces, qui proviennent des journaux d'audit des systèmes de surveillance, sont des données dites de connexion (DDC) et ne sont pas considérées comme des DCP. À cet égard, plusieurs aspects sont à considérer.

La loi encadre la conservation et la consultation des DDC en réservant aux opérateurs télécom l'usage de ces données dans certaines conditions comme la sécurité de leurs réseaux ou encore leur besoin de facturation. Certains d'entre eux en font un usage commercial cependant. Ainsi l'opérateur AT&T a-t-il développé son système *Time Warner* afin de croiser les métadonnées de connexion de ses utilisateurs : messages, heure d'émission ou destinataire. Ces DDC sont aussi précieuses pour les forces de l'ordre américaines afin d'obtenir des informations sur la géolocalisation de suspects.

L'exploitation de ces traces à l'aide des outils de Big Data permet d'élever la connaissance sur les comportements malveillants en vue de les anticiper, un atout pour la lutte contre le terrorisme. En revanche, dans le cas d'usage de techniques d'anonymisation, les experts de sécurité seront dans l'impossibilité d'exploiter ces traces numériques et par là-même, d'établir une responsabilité ou de capitaliser sur ces données.

Par Véronique Legrand,
Professeure du Cnam,
chaire Sécurité informatique,
membre du Centre d'études et de recherche en informatique et communications.

Droit Informatique Numérique Sécurité

Le dernier Cnam mag'

LE CNAM MAG' #9

Société numérique, société inclusive ?

1 mai 2018

+ Retrouvez tous les numéros

Suivez ses enseignements

L'unité d'enseignement Cybersécurité : référentiel, objectifs et déploiement

http://blog.cnam.fr/technologie/les-big-data/big-data-et-cybersecurite-quelles-armes-face-a-cette-ruee-vers-ce-nouvel-o