« Les difficultés du procédé d'anonymisation des données »

Intervention du 20 mai 2022 dans le cadre du 2e Printemps de la donnée UBFC par Mickael Taubaty et Nicolas Kern, étudiants en Master Droit du numérique, UFR SJEPG – Université de Franche-Comté.

En collaboration avec Delphine Martin, MCF de droit privé, UFR SJEPG, Université de Franche-Comté.

Introduction :

Selon la loi Informatique et Libertés du 6 janvier 1978, une donnée personnelle est définie « comme toute information se rapportant directement ou indirectement à une personne physique identifiée ou identifiable »[1]. Une seule donnée peut suffire pour identifier une personne par exemple : son nom ou son prénom. Mais il peut aussi s'agir aussi de plusieurs données indirectes comme la composition du foyer, la commune d'habitation, une adresse IP…permettant de retrouver plus ou moins facilement la personne concernée.

Le traitement de ces données personnelles est encadré par le Règlement européen sur la protection des données (RGPD) entré en vigueur le 25 mai 2018[2]. Par traitement, il faut entendre toute opération qui peut être effectuée sur des données personnelles comme une simple modification, une opération de stockage, ou encore un partage de données.

Deux textes principaux réglementent le traitement des données personnelles : la loi informatique et liberté de 1978 et le Règlement général sur la protection des données (RGPD) applicable à toute collecte de données personnelles réalisée au sein de l'Union européenne.

Le RGPD pose un certain nombre de principes relatifs à la collecte et au traitement des données personnelles à l'article 5 :

- le principe de finalité : le traitement des données personnelles doit être réalisé dans un but bien précis, légal et légitime.

- le principe de proportionnalité, de pertinence et de minimisation : les données concernées et conservées doivent présenter un intérêt pertinent et strictement nécessaire à la finalité du traitement.

[1] Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés dite loi LIL.
[2] Règlement (UE) 2016/679 du 27 avril 2016 sur la protection des données.

- le principe de sécurisation et de confidentialité-: le responsable de traitement doit garantir la sécurité et la confidentialité des informations collectées qu'il a en sa possession et en limiter l'accès.

- le principe de conservation limitée des données : les données ne peuvent pas être conservées indéfiniment sauf si l'on procède à leur anonymisation. La durée est fixée préalablement en fonction de la nature des données traitées et de la finalité de traitement soit par un texte légal, soit par le responsable de traitement.

Exemple 1 : les données d'un candidat qui postule dans une entreprise pourront être gardées pendant 2 ans maximum par le service des Ressources humaines avant leur suppression définitive

Exemple 2 : l'article L3243-4 du Code du travail impose à l'employeur de conserver un double du bulletin de paie du salarié pendant 5 ans maximum.

Exemple 3 : les données de facturation sont conservées pendant 10 ans pour faciliter une éventuelle coopération avec les autorités judiciaires.

Lorsque les obligations prévues par le RGPD ne sont pas respectées par le responsable de traitement, les sanctions sont lourdes : pour une entreprise, jusqu'à 4% du CA mondial de l'exercice précédent ou jusqu'à 20 millions d'€ d'amende administrative[3].

Cependant, les règles posées par le RGPD relatives au traitement des données peuvent être écartées au moyen de la technique de l'anonymisation des données. Ce procédé élimine, en effet, toute possibilité de ré-identification des personnes concernées par la collecte et donc permet de traiter leurs données dans le respect de leurs droits et libertés. Toutefois, ce procédé constitue en lui-même un traitement ultérieur de données personnelles et doit donc satisfaire à l'exigence de compatibilité au regard des motifs juridiques et des circonstances du traitement ultérieur (peut-être rappeler les hypothèses légales de traitement ultérieur, statistiques, archivistiques, scientifiques). Il faut également garder à l'esprit que si les données traitées sortent du champ d'application du RGPD elles restent exposées à d'autres dispositions (comme celles relatives à la confidentialité des communications – secret des correspondances).

Le procédé de l'anonymisation n'est pas obligatoire, ce n'est qu'un moyen de parvenir à garantir le respect des droits des personnes concernées (ce qui est l'objectif premier du RGPD et de la loi LIL). Toutefois, par exception, pour certains documents administratifs et en fonction de la nature des données qu'ils contiennent, l'anonymisation préalable est obligatoire pour les diffuser en ligne, ceci en application du Code des relations entre le public et l'administration. A défaut le consentement des personnes concernées est nécessaire[4]. Attention également à l'existence de

[3] Art. 83 RGPD.
[4] Art. L-312-1-1 CRPA.

partenariats avec les administrations, certaines informations administratives ne peuvent être diffusées sans anonymisation préalable

Avant de définir l'anonymisation, il convient de la distinguer de la pseudonymisation. S'il est plus fastidieux d'anonymiser des données personnelles que de les pseudonymiser, la première technique offre plus de garanties au niveau de la protection des droits et libertés des personnes que la seconde.

Selon la CNIL, la pseudonymisation est "un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données à une personne physique identifiée sans information supplémentaire"[5]. Autrement dit, cette méthode vise à réaliser le traitement de données personnelles de manière à ce qu'on ne puisse plus associer les données traitées à une personnes en particulier sans devoir rechercher d'autres informations supplémentaires. Cela se fera par un remplacement des données permettant d'identifier directement une personne (nom, prénom, etc.) par d'autres données (alias, numéros séquentiel (suite ordonnée de symboles), etc.). Ce qui caractérise la pseudonymisation c'est le caractère réversible de l'opération. Trouver l'identité de la personne reste possible avec des informations supplémentaires le but n'est donc pas de rendre toute réidentification impossible.

-Exemple pratique présent sur le site de la Cnil : un économiste qui, dans le cadre de ses travaux de recherche, forme un partenariat avec la CAF. La CAF transmet au chercheur les données concernant le montant des allocations allouées en remplaçant les nom et dates de naissances par un identifiant unique + remplacer les adresses par le nom des communes. Les données concernant la composition du foyer restent les mêmes de sorte que dans les petites communes, ces infos pourraient permettre de retrouver les personnes concernées. Si, dans le tableau transmis par la CAF il est précisé que 100% des personnes de plus de 90 ans vivant seul bénéficie d'une aide au logement. Il serait tout à fait envisageable de retrouver l'identité de la personne en question avec les informations de la commune ciblée.

Les données résultant d'une pseudonymisation sont intégralement soumises aux obligations du RGPD donc cela implique le respect du principe de conservation, du respect des droits des personnes concernées, la mise en œuvre de modalités techniques pour les sécuriser, etc... La pseudonymisation n'est pas obligatoire, mais recommandée car elle contribue à sécuriser les données et à garantir un traitement des données personnelles conforme au Règlement européen.

La CNIL définit l'anonymisation comme « un traitement de données personnelles qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute réidentification de la personne, par quelque moyen que ce soit »[6]. Il s'agit, contrairement à la pseudonymisation d'un procédé irréversible**.** Ce n'est donc pas une simple action visant à remplacer les données**.** L'anonymisation entraine une destruction de certaines données**.** Ainsi,

[5] Avis du 19 mai 2020, « L'anonymisation des données personnelles ».
[6] Avis *préc.*

si l'anonymisation est bien réalisée, les données concernées ne sont plus soumises à la réglementation RGPD car elles ne représentent plus aucun risque pour la vie privée des personnes concernées.

-Exemple : établir un tableau des consommations d'énergie des foyers présent dans un quartier sur une période de 3 mois. Si on ne fait que préciser que 20% de l'électricité total est utilisé la nuit et que 80% des foyers consommes + en heures creuses = aucun moyen de remonter les informations pour identifier les foyers en question. Les données ainsi décrite sont moins précises, moins fines, mais elles permettent une anonymisation satisfaisante.

Il est pratique ou nécessaire de recourir à l'anonymisation lors de l'utilisation massive ou de la recherche massive de données, mais il convient de rester vigilant quant à la manière d'anonymiser les données car si un ensemble de données publié en ligne est présenté comme anonymisé alors qu'il contient des données personnelles identifiantes, il s'agira d'une violation de données soumise aux sanctions prévues par le Règlement européen.

Lorsque la violation de données est caractérisée, des procédures sont à effectuer comme la mise en place de mesures visant à détecter immédiatement une violation, à l'endiguer rapidement, à analyser les risques engendrés par l'incident et à déterminer s'il convient de la notifier à l'autorité de contrôle, voire aux personnes concernées.

En pratique :

En pratique, La CNIL recommande de se poser les bonnes questions et d'avoir les bons réflexes :
- Supprimer les données et valeurs d'identification directe permettant une réidentification ex : âge précis.

- Distinguer les données importantes et non-essentielles et ne conserver que les données utiles aux recherches dans le respect du principe de minimisation des données.

- Définir la finalité des données et les priorités.

Cette démarche permet de faciliter l'anonymisation.

Il existe plusieurs procédés d'anonymisations, à titre d'exemples seront cités les plus courants :

- **Randomisation ou tirage au sort** : procédé selon lequel l'attribution d'un traitement à une personne sera réalisée de façon aléatoire. Autrement dit, modifier les attributs dans une sélection de données pour les rendre moins précises afin de les protéger du risque d'interférence**.** Permuter des informations pour altérer la véracité des informations non nécessaires à la compréhension de la finalité du traitement. Ce procédé est généralement utilisé dans le cadre de la médecine lorsque l'on veut par exemple mesurer les effets d'un nouveau traitement sur une partie de la population. Elle peut aussi être utilisée afin de classifier aléatoirement des groupes de personnes concernées par un traitement de données à caractère personnelles.

- ***Généralisation*** : généraliser les données en modifiant leur échelle ou ordre de grandeur. Ex : si je dois faire des statistiques sur le nombre de maisons possédant une piscine et une cheminée, il sera plus difficile de retrouver ces maisons si j'effectue mes recherches à l'échelle d'une région plutôt qu'à l'échelle d'un quartier.

Comment vérifier l'efficacité de cette anonymisation ?

Des préconisations ont été formulées dans un avis du 10 avril 2014 par le G29, devenu le CEPD (Comité Européen de la Protection des Données)[7].

Trois critères permettent de s'assurer qu'un jeu de données est véritablement anonyme :

- La non-individualisation : un individu ne doit pas pouvoir être isolé dans un jeu de données.
- La non-corrélation : des ensembles de données distincts concernant un même individu ne doivent pas pouvoir être reliés entre eux.
- La non-inférence : de nouvelles informations relatives à un individu ne doivent pas pouvoir être déduites.

Lorsque ces critères ne sont pas réunis, le chercheur doit pouvoir démontrer que le risque de réidentification par des moyens raisonnables est nul.

L'une des difficultés est de s'adapter à l'évolution des moyens et des connaissances techniques et technologiques afin de maintenir cette anonymisation ce qui implique d'effectuer une veille sur le plan technique notamment. L'article 26 du RGPD énonce, en effet, que l'on tient « compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci. ». De plus, il faut pouvoir maintenir cette anonymisation pour l'avenir donc veiller à ce que les évolutions techniques ne permettent pas une réidentification qui était impossible au moment de sa mise en place.

Les procédés techniques :

Afin d'appréhender l'anonymisation de manière plus concrète, il est plus simple de l'illustrer par des procédés techniques. Cependant il est important de souligner qu'actuellement aucun procédé ne permet d'anonymiser des données de manière infaillible, notamment en raison de l'évolution permanente des techniques d'attaque. De plus, certaines méthodes sont critiquées en raison de leur obsolescence et donc de leur caractère faillible.

Pour contrer ce risque d'attaque, il est souvent utilisé cumulativement plusieurs techniques. Aujourd'hui la question n'est plus « est ce que l'on va se faire attaquer ? » mais « quand va-t-on se faire attaquer ? »

Ces techniques sont très nombreuses, seront seuls citées les plus classiques et les plus utilisées.

[7] Avis 05/2014 sur les Techniques d'anonymisation, adopté le 10 avril 2014,

- **Le « k-anonymat »** qui est un standard de l'industrie (notamment utilisé par Google), cela désigne une technique permettant de masquer l'identité d'individus dans un groupe de personnes semblables.

Cependant, on peut lui reprocher de ne poser aucune contrainte sur les valeurs sensibles.

C'est pour cela que des experts ont mis en place une technique souvent utilisée comme un complément au « k-anonymat », il s'agit de la technique dite « l-diversité » qui permet une meilleure anonymisation en examinant s'il y a assez de valeurs différentes à l'intérieur de chaque groupe de personnes semblables pour empêcher la déduction par homogénéité des données d'une personne. Pour simplifier, elle permet d'éviter que les données collectées par une entreprise auprès d'un groupe de personnes ne permettent de les identifier.

- La « permutation » : ce processus vise à garder les attributs exacts de chaque donnée, mais à les attribuer de manière aléatoire à d'autres individus.

Toutefois, le risque de remonter aux personnes à partir de ces données si les attaquants ont connaissance de quasi-identifiants.

Pour limiter en partie les risques de ré-identification, il a été développé récemment des nouvelles techniques qui reposent sur des algorithmes, comme le concept de *differential privacy* qui permet de s'abstraire complètement des connaissances que peut avoir l'attaquant en générant des fausses données qui sont « compensées » avec les vraies données.

Il est possible d'aller plus loin avec le « calcul sécurisé multipartite » qui relève de la cryptographie lorsque la personne qui anonymise les données n'a pas confiance dans ces prestataires de service par exemple. C'est une procédure plus longue et plus complexe.

Parmi les procédés d'anonymisation, certains sont recommandés par la CNIL :

Exemple : mise en place en 2020 par Webdata d'une méthode qui consiste à proposer une solution basée sur des avatars pour anonymiser les dossiers de patients à l'hôpital. Ces données peuvent être ré-utilisées dans le domaine de la recherche médicale sans risque pour les malades d'être ré-identifiés.
De manière plus technique, l'algorithme chargé d'anonymiser les données permet de créer un profil à partir des données d'un individu différent du patient, dans ce cas l'intégralité des données sont modifiées.

Les difficultés techniques de l'anonymisation :

S'agissante des difficultés plus techniques, il convient de souligner qu'il est impossible de se protéger totalement des personnes ayant des informations sur les identifiants et qu'aucun système n'est véritablement fiable.

Et le problème le plus courant est que les attaques perpétrées sur les données anonymisées passent souvent par une corrélation avec des données publiques insoupçonnées, c'est-à-dire que certaines données en ligne permettent de recouper des données anonymisées pour obtenir des quasis identifiants.

Deux affaires illustrent les dangers actuels de l'accès à des données insuffisamment anonymisées :

- Affaire Commission Taxi & Limousine de New York (NYC) : il s'agit d'une affaire où une société donne libre accès, à cause de la loi américaine FOIL, à un jeu de données de 20 Go contenant plus de 173 millions de courses, avec notamment lieu et date de début et fin de la course, ainsi que le numéro de licence du conducteur et l'identifiant du taxi anonymisé mais des attaquants ont par une technique nommée attaque par force brute réussi à réidentifier toutes les voitures et les conducteurs.
- Cas d'une journaliste allemande qui s'est présentée comme dirigeante d'une start-up et qui à ce titre a obtenu l'accès à un « jeu d'essai » de l'historique complet de navigation de 3 millions d'utilisateurs allemands anonymisés pendant un mois auprès d'une société de data brokerage (courtier en donnée) sous prétexte d'acheter ensuite les données. Cette journaliste a pu, avec très peu d'éléments retrouver une personne parmi ces utilisateurs.

Comment s'assurer dès lors et abstraction faite des critères dégagés par le G29 de l'efficacité de l'anonymisation ?

Il est d'abord nécessaire de définir à chaque fois une démarche d'anonymisation adaptée à chaque ensemble de données afin d'identifier la combinaison de techniques qui répondent le mieux à l'objectif recherché.

C'est d'ailleurs l'individualisation que préconise le G29 dans son avis de 2014 : "Une solution d'anonymisation doit être construite au cas par cas et adaptée aux usages prévus".

Certaines solutions peuvent être préconisées pour renforcer l'efficacité de l'anonymisation comme le fait de travailler sur des groupes d'individus et non plus sur des individus uniques. Toutefois, cette solution peut ne pas convenir aux entreprises qui souhaitent conserver des données les plus précises possibles.

La solution la plus efficace reste sans doute celle de *« Se mettre dans la peau de l'attaquant »* du fait que les attaques sont orchestrées par des personnes qui ont pour but de pénétrer un système, il est donc judicieux de simuler des attaques afin de d'en identifier les failles pour mieux anticiper des attaques futures.

Par ailleurs, le RGPD recommande la création de certifications attestant la conformité des entreprises afin d'assurer et de garantir que l'anonymisation est efficace et donc d'éviter une violation de données.

A ce sujet, la CNIL recommande deux types de certifications :

-Soit une certification des solutions d'anonymisation. Dans cette hypothèse c'est une solution pour répondre aux exigences techniques afin d'anonymiser des données qui est l'objet de la certification. Mais il n'y a aucune certitude sur l'application de cette solution par l'entreprise certifiée.

-Soit une certification des entreprises, donc des systèmes et processus mis en place pour garantir l'anonymité des données (plus conformes aux préconisations du RGPD) mais cela n'exclut pas la possibilité d'une violation des données en cas de faille avérée.

Conclusion :

Le fait de procéder à une anonymisation complète des données permet d'échapper entièrement à l'application des règles du RGPD. L'anonymisation rend impossible toute réidentification des personnes concernées par le traitement, il n'y a donc plus rien de potentiellement attentatoire à leurs libertés, il n'est donc plus nécessaire de respecter les principes applicables à la collecte et au traitement des données qu'elles soient sensibles ou non. C'est là l'enjeu premier de cette pratique. Elle se distingue nettement de la pseudonymisation, car même si l'opération peut paraître plus fastidieuse, le résultat reste hautement bénéfique, car il sera ensuite possible de conserver et de réutiliser les données sans aucune restriction.

De plus, il est nécessaire de rappeler que l'anonymisation représente un effort constant. Ce n'est pas parce que l'on va anonymiser une fois les données que le travail sera terminé. Comme dit précédemment les techniques et les technologies peuvent évoluer permettant alors une réidentification là où elle était censée être impossible lors de la première anonymisation (exemple : des données anonymisées dans le cadre de la recherche comme des statistiques peuvent servir à étoffer des profils existants et donc générer de nouveaux problèmes). Donc il y a une très grande liberté en ce qui concerne le traitement et la diffusion des données anonymisées mais à condition que l'anonymisation reste efficace. Il convient donc de procéder à une veille juridique et technologique en se renseignant sur l'actualité, mais aussi auprès des juristes qui ont des compétences RGPD (comme le DPO). L'anonymisation s'inscrit dans le temps et implique que le responsable de traitement réévalue régulièrement les risques associés au traitement, ce n'est pas une opération unique et perpétuelle.

Sources :

Définition anonymisation CNIL : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

Cnil sur la difficulté : https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

L'anonymisation des données, un traitement clé pour l'open data : https://www.cnil.fr/fr/lanonymisation-des-donnees-un-traitement-cle-pour-lopen-data

Difficulté technique : https://medium.com/meetech/de-la-difficult%C3%A9-technique-de-lanonymisation-ou-comment-mal-anonymiser-ses-donn%C3%A9es-b1cc44f623cc

Considérant 26 RGPD : https://gdpr-text.com/fr/read/recital-26/

L'enjeu de l'anonymisation a l'heure du big data : https://www.cairn.info/revue-francaise-des-affaires-sociales-2017-4-page-79.htm?try_download=1

Est-il vraiment possible d'anonymiser les données sensibles ? : https://www.journaldunet.com/solutions/dsi/1459133-est-il-vraiment-possible-d-anonymiser-les-donnees-sensibles/

PDF : https://tel.archives-ouvertes.fr/tel-01783967/document

Anonymiser pour alléger les contraintes liées aux données personnelles : https://www.usine-digitale.fr/article/anonymiser-pour-alleger-les-contraintes-liees-aux-donnees-personnelles.N830905

Article recommandation Europe : https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_fr.pdf

Article sur l'obligation d'anonymiser les personnes : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000033205514/

Exemple PMSI : https://controverses.minesparis.psl.eu/public/promo16/promo16_G13/www.controverses-minesparistech-3.fr/_groupe13/anonymisation-ou-pseudonymisation-des-donnees-de-sante/index.html

Wedata : https://www.usine-digitale.fr/article/la-cnil-approuve-wedata-pour-l-anonymisation-des-donnees-de-sante.N1024644

Donnée de santé: https://www.datanaos.com/anonymisation_des_donnees_de_sante

Techniques : https://www.didaktic.fr/fair-open-data/structures-et-groupes-de-travail/dispositifs-techniques-danonymisation-et-pseudonymisation-des-donnees-de-sante/

Des bases légales utiles :
https://www.cnil.fr/fr/recherche-scientifique-hors-sante-quelle-base-legale-pour-un-traitement-de-recherche

Pour prévoir les questions : durée de conservation des données dans le cadre de la recherche + comment conserver les données dans le cadre d'une recherche :
https://www.cnil.fr/fr/recherche-scientifique-hors-sante/durees-conservations-donnees

https://medium.com/meetech/lanonymisation-de-donn%C3%A9es-une-qu%C3%AAte-encore-difficile-au-lendemain-de-l-entr%C3%A9e-en-vigueur-du-rgpd-6d25522fd2d6

Anonymisation des décisions de justice : https://www.macsf.fr/responsabilite-professionnelle/cadre-juridique/anonymisation-des-decisions-de-justice#:~:text=La%20pseudonymisation%20n'est%20qu,des%20parties%20reste%20souvent%20possible.