RGPD

À l'ère de l'IA et du scraping, la fin des « données anonymes » ?

Dans un arrêt rendu le 26 avril 2023 par le tribunal de l'Union européenne (affaire T-557/20), le juge européen est revenu sur des notions très contestées de « données personnelles » et « données anonymes », ainsi que sur les critères permettant de les distinguer.

Ces notions, qui peuvent paraître évidentes, font en réalité l'objet de bon nombre de débats, du fait des conséquences qui découlent de leurs statuts ; en effet, la qualification d'une information en « *donnée personnelle* » signifie que le RGPD et les obligations qui en découlent (obligation d'information, consentement, mise en place de mesures techniques et organisationnelles...) s'appliquent, alors qu'à l'inverse, une « *donnée anonyme* », sort du champ du RGPD, et s'affranchit de ces obligations. Les critères permettant de déterminer si une donnée est anonyme ou non ont donc une grande importance, puisque c'est tout le champ de protection du RGPD qui est en jeu.

Le juge européen pose ici un critère de distinction intéressant, mais qui ne nous semble pas prendre en compte les nouvelles techniques, notamment de scrapping ou d'intelligence artificielle, et qui mène à se poser la question suivante : existe-t-il vraiment, aujourd'hui, des données anonymisées ou anonymes ?

Distinction entre une donnée pseudonymisée et une donnée anonymisée : le critère posé par le tribunal de l'UE

Dans l'arrêt précité, le tribunal de l'UE était amené à se prononcer dans une affaire opposant le Contrôleur européen de la protection des données (CEPD) et le « *Conseil de Résolution Unique* » (CRU), institution européenne, sur le fondement non pas du RGPD, mais du règlement n° 2018/1725 du 23 octobre 2018, spécifiquement dédié au traitement des données à caractère personnel par les institutions, organes et organismes de l'Union européenne.

Dans le cadre du traitement de demandes de dédommagements, le CRU avait ainsi mis en place une procédure en deux temps, avec une collecte des données personnelles des personnes concernées dans un premier temps, et dans un second temps la collecte et le traitement des commentaires de ces dernières, sous format pseudonymisé.
Dans le cadre de cette seconde phase, le cabinet de conseil privé Deloitte est intervenu en tant que sous-traitant, et a donc reçu ces commentaires de la part du CRU.

Il a été reproché au CRU de ne pas avoir informé les personnes concernées de la présence de ce destinataire, à qui des « *données personnelles* » avaient été transmises selon le CEPD. Le CEPD a en effet rendu dans ce cadre un avis considérant que les commentaires livrés au cabinet Deloitte bien que « *pseudonymisés* », étaient des données personnelles, et non des « *données anonymisées* ».

Pour poser ses critères de distinction entre une donnée personnelle et une donnée anonyme, le tribunal de l'UE rappelle tout d'abord la logique de l'arrêt Breyer, rendu par la CJUE le 19 octobre 2016 (C-582/14) au sujet de l'adresse IP : une adresse IP peut constituer une donnée personnelle à l'égard du fournisseur de services de médias en ligne qui l'enregistre si ce dernier dispose de moyens légaux pour faire identifier la personne concernée grâce aux informations supplémentaires dont dispose le fournisseur d'accès à internet de cette personne.

Le tribunal de l'UE en a déduit que pour qualifier une donnée d'anonyme ou de personnelle, il convenait de se demander si le(s) destinataire(s) de cette donnée avai(en)t la possibilité d'accéder à des informations supplémentaires permettant de combiner et de « *réidentifier de manière raisonnable les personnes concernées par ces données transmises* ». Sur la base de ce critère, il a été conclu que les données transmises au cabinet Deloitte étaient « *anonymisées* », car ce dernier n'avait pas la possibilité de ré-identifier de manière raisonnable les personnes concernées.

Ainsi, il est possible, selon cet arrêt, qu'une donnée soit considérée comme une « *donnée personnelle* » entre les mains d'une entité et comme une donnée « *anonyme* » entre les mains d'une autre. Cette distinction devra être opérée par le responsable de traitement qui devra au cas par cas et selon le contexte, s'assurer si oui ou non les destinataires de ses données disposent ou non des « *moyens raisonnables* » permettant d'accéder à des informations supplémentaires qui, combinées aux données transmises, permettraient de réidentifier la personne concernée.

À l'ère de l'intelligence artificielle et du scraping, la collecte de données de masse sur internet prend de l'ampleur : la fin de la donnée anonyme ?

L'anonymisation n'est pas définie dans le règlement UE n°2016/679 dit RGPD, mais cette mesure technique est recherchée par un grand nombre d'acteurs du numérique, qui voit dans cette dernière la possibilité infinie d'utiliser des jeux de données, sans avoir à se conformer aux règles jugées contraignantes du RGPD.

La CNIL en donne la définition suivante : « *l'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.* »

Au niveau européen, trois critères sont pris en compte par les autorités de protection pour s'assurer qu'une donnée est bien « *anonyme* » : (1) l'individualisation (l'impossibilité d'isoler un individu dans un jeu de donnée), (2) la corrélation (l'impossibilité de combiner des jeux de données distincts concernant une même personne), et (3) l'inférence (l'impossibilité de déduire de façon certaine de nouvelles informations sur l'individu).Un avis de l'ancien CEPD (« *G29* »), bien qu'ancien – 2014, dédie par ailleurs des lignes directrices aux différentes techniques d'anonymisation[1].

Mais cette technique de l'anonymisation est-elle adaptée aux techniques actuelles de collectes massives de données publiques indifférenciées dont ont recours un grand nombre d'acteurs aujourd'hui ? C'est notamment l'essence même du scraping et de l'intelligence artificielle.

Si l'on transpose les critères posés par le tribunal de l'UE au monde du scraping et de l'intelligence artificielle, ces derniers semblent difficiles à appliquer, pour plusieurs raisons. D'une part les responsables de traitement dans ces cas-là n'ont pas toujours conscience des données qu'ils (ou leurs sous-traitants) récoltent, d'autre part ils n'ont pas non plus le contrôle sur toutes les potentielles combinaisons possibles entre ces différentes données récoltées et réalisées par des algorithmes, enfin, le but de ces techniques innovantes est de s'alimenter en continu de nouvelles données trouvées en ligne.

Comment dans ce cas sera-t-il possible pour ces derniers de s'assurer qu'une donnée initialement récoltée n'est pas susceptible d'être combinée avec d'autres données pour « *réidentifier de manière raisonnable* » un individu ?

En se fondant sur ce critère-là, il nous paraît donc aujourd'hui difficile d'estimer qu'il existe des « *données anonymes* » à l'ère de l'intelligence artificielle et du scraping.

Ce constat n'est pas dramatique ni fatal, mais il doit encourager les concepteurs de ces innovations à toujours garder le réflexe de la protection des données des citoyens au cœur de leur business, sans se retrancher derrière un possible « *anonymat* ».

Alexandra ITEANU
Avocat à la Cour
Iteanu société d'avocats

Notes

(1) Avis 05/2014 sur les techniques d'anonymisation

Vous avez envie de vous exprimer sur un sujet qui vous tient à cœur, de partager votre analyse avec la communauté des lecteurs d'Expertises, d'exposer un point de vue différent sur un article déjà publié, de lancer un débat sur un thème émergent, ou simplement de commenter l'actualité du droit du numérique ?

Contactez la rédactrice en chef d'Expertises Sylvie Rozenfeld sr@expertises.info

RGPD

À l'ère de l'IA et du scraping, la fin des « données anonymes » ?

Dans un arrêt rendu le 26 avril 2023 par le tribunal de l'Union européenne (affaire T-557/20), le juge européen est revenu sur des notions très contestées de « données personnelles » et « données anonymes », ainsi que sur les critères permettant de les distinguer.

Ces notions, qui peuvent paraître évidentes, font en réalité l'objet de bon nombre de débats, du fait des conséquences qui découlent de leurs statuts ; en effet, la qualification d'une information en « *donnée personnelle* » signifie que le RGPD et les obligations qui en découlent (obligation d'information, consentement, mise en place de mesures techniques et organisationnelles...) s'appliquent, alors qu'à l'inverse, une « *donnée anonyme* », sort du champ du RGPD, et s'affranchit de ces obligations. Les critères permettant de déterminer si une donnée est anonyme ou non ont donc une grande importance, puisque c'est tout le champ de protection du RGPD qui est en jeu.

Le juge européen pose ici un critère de distinction intéressant, mais qui ne nous semble pas prendre en compte les nouvelles techniques, notamment de scrapping ou d'intelligence artificielle, et qui mène à se poser la question suivante : existe-t-il vraiment, aujourd'hui, des données anonymisées ou anonymes ?

Distinction entre une donnée pseudonymisée et une donnée anonymisée : le critère posé par le tribunal de l'UE

Dans l'arrêt précité, le tribunal de l'UE était amené à se prononcer dans une affaire opposant le Contrôleur européen de la protection des données (CEPD) et le « *Conseil de Résolution Unique* » (CRU), institution européenne, sur le fondement non pas du RGPD, mais du règlement n° 2018/1725 du 23 octobre 2018, spécifiquement dédié au traitement des données à caractère personnel par les institutions, organes et organismes de l'Union européenne.

Dans le cadre du traitement de demandes de dédommagements, le CRU avait ainsi mis en place une procédure en deux temps, avec une collecte des données personnelles des personnes concernées dans un premier temps, et dans un second temps la collecte et le traitement des commentaires de ces dernières, sous format pseudonymisé.
Dans le cadre de cette seconde phase, le cabinet de conseil privé Deloitte est intervenu en tant que sous-traitant, et a donc reçu ces commentaires de la part du CRU.

Il a été reproché au CRU de ne pas avoir informé les personnes concernées de la présence de ce destinataire, à qui des « *données personnelles* » avaient été transmises selon le CEPD. Le CEPD a en effet rendu dans ce cadre un avis considérant que les commentaires livrés au cabinet Deloitte bien que « *pseudonymisés* », étaient des données personnelles, et non des « *données anonymisées* ».

Pour poser ses critères de distinction entre une donnée personnelle et une donnée anonyme, le tribunal de l'UE rappelle tout d'abord la logique de l'arrêt Breyer, rendu par la CJUE le 19 octobre 2016 (C-582/14) au sujet de l'adresse IP : une adresse IP peut constituer une donnée personnelle à l'égard du fournisseur de services de médias en ligne qui l'enregistre si ce dernier dispose de moyens légaux pour faire identifier la personne concernée grâce aux informations supplémentaires dont dispose le fournisseur d'accès à internet de cette personne.

Le tribunal de l'UE en a déduit que pour qualifier une donnée d'anonyme ou de personnelle, il convenait de se demander si le(s) destinataire(s) de cette donnée avai(en)t la possibilité d'accéder à des informations supplémentaires permettant de combiner et de « *réidentifier de manière raisonnable les personnes concernées par ces données transmises* ». Sur la base de ce critère, il a été conclu que les données transmises au cabinet Deloitte étaient « *anonymisées* », car ce dernier n'avait pas la possibilité de ré-identifier de manière raisonnable les personnes concernées.

Ainsi, il est possible, selon cet arrêt, qu'une donnée soit considérée comme une « *donnée personnelle* » entre les mains d'une entité et comme une donnée « *anonyme* » entre les mains d'une autre. Cette distinction devra être opérée par le responsable de traitement qui devra au cas par cas et selon le contexte, s'assurer si oui ou non les destinataires de ses données disposent ou non des « *moyens raisonnables* » permettant d'accéder à des informations supplémentaires qui, combinées aux données transmises, permettraient de réidentifier la personne concernée.

À l'ère de l'intelligence artificielle et du scraping, la collecte de données de masse sur internet prend de l'ampleur : la fin de la donnée anonyme ?

L'anonymisation n'est pas définie dans le règlement UE n°2016/679 dit RGPD, mais cette mesure technique est recherchée par un grand nombre d'acteurs du numérique, qui voit dans cette dernière la possibilité infinie d'utiliser des jeux de données, sans avoir à se conformer aux règles jugées contraignantes du RGPD.

La CNIL en donne la définition suivante : « *l'anonymisation est un traitement qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne par quelque moyen que ce soit et de manière irréversible.* »

Au niveau européen, trois critères sont pris en compte par les autorités de protection pour s'assurer qu'une donnée est bien « *anonyme* » : (1) l'individualisation (l'impossibilité d'isoler un individu dans un jeu de donnée), (2) la corrélation (l'impossibilité de combiner des jeux de données distincts concernant une même personne), et (3) l'inférence (l'impossibilité de déduire de façon certaine de nouvelles informations sur l'individu).Un avis de l'ancien CEPD (« *G29* »), bien qu'ancien – 2014, dédie par ailleurs des lignes directrices aux différentes techniques d'anonymisation[1].

Mais cette technique de l'anonymisation est-elle adaptée aux techniques actuelles de collectes massives de données publiques indifférenciées dont ont recours un grand nombre d'acteurs aujourd'hui ? C'est notamment l'essence même du scraping et de l'intelligence artificielle.

Si l'on transpose les critères posés par le tribunal de l'UE au monde du scraping et de l'intelligence artificielle, ces derniers semblent difficiles à appliquer, pour plusieurs raisons. D'une part les responsables de traitement dans ces cas-là n'ont pas toujours conscience des données qu'ils (ou leurs sous-traitants) récoltent, d'autre part ils n'ont pas non plus le contrôle sur toutes les potentielles combinaisons possibles entre ces différentes données récoltées et réalisées par des algorithmes, enfin, le but de ces techniques innovantes est de s'alimenter en continu de nouvelles données trouvées en ligne.

Comment dans ce cas sera-t-il possible pour ces derniers de s'assurer qu'une donnée initialement récoltée n'est pas susceptible d'être combinée avec d'autres données pour « *réidentifier de manière raisonnable* » un individu ?

En se fondant sur ce critère-là, il nous paraît donc aujourd'hui difficile d'estimer qu'il existe des « *données anonymes* » à l'ère de l'intelligence artificielle et du scraping.

Ce constat n'est pas dramatique ni fatal, mais il doit encourager les concepteurs de ces innovations à toujours garder le réflexe de la protection des données des citoyens au cœur de leur business, sans se retrancher derrière un possible « *anonymat* ».

Alexandra ITEANU
Avocat à la Cour
Iteanu société d'avocats

Notes

(1) Avis 05/2014 sur les techniques d'anonymisation

Vous avez envie de vous exprimer sur un sujet qui vous tient à cœur, de partager votre analyse avec la communauté des lecteurs d'Expertises, d'exposer un point de vue différent sur un article déjà publié, de lancer un débat sur un thème émergent, ou simplement de commenter l'actualité du droit du numérique ?

Contactez la rédactrice en chef d'Expertises Sylvie Rozenfeld sr@expertises.info