Explicatif : Anonymisation des données

L'un des principaux objectifs d'un cadre de gouvernance des données est la gestion des questions de protection de la vie privée soulevées par ces données. La plupart des données en lien avec l'activité humaine sont de nature personnelle, visant des individus identifiables. Ces données personnelles sont protégées par les diverses lois sur les renseignements personnels qui ont cours un peu partout.

L'une des solutions souvent proposées pour utiliser des données personnelles sans risque pour la vie privée est d'en retirer tout élément d'information pouvant directement ou indirectement identifier un individu en particulier. C'est ce que l'on appelle l'anonymisation. Les identifiants directs comprennent tout renseignement pouvant identifier, seul ou en combinaison avec d'autres renseignements, un individu en particulier : nom, numéro de téléphone, adresse, numéro de carte de crédit, numéro d'assurance sociale, etc. Pour leur part, les identifiants indirects comprennent tout renseignement pouvant servir à identifier un individu en particulier, mais uniquement lorsqu'il est combiné à un autre renseignement, notamment ceux mettant l'individu en contexte.

L'anonymisation des données suppose inévitablement un compromis entre la vie privée et le niveau de détail. Plus une catégorie de données est détaillée, plus elle contient d'information, y compris des renseignements personnels. Pour sa part, le respect de la vie privée exige de minimiser la communication de renseignements personnels, limitant d'autant l'utilité de la catégorie de données ainsi dépersonnalisée. La protection de la vie privée exige fondamentalement qu'un individu ait le droit et les moyens de contrôler ses renseignements personnels. Ainsi, un individu pourrait décider de refuser que ses renseignements personnels soient utilisés à des fins qu'il désapprouve, et ce, même si ses renseignements ont été dépersonnalisés. Les droits juridiques liés à la protection des renseignements personnels s'appliquent tout autant à des renseignements dépersonnalisés, notamment pour ce qui est de la spécification des fins, de la reddition de compte et de la transparence. Il faut se rappeler qu'un individu reconnaîtra toujours un renseignement le concernant, qu'il puisse l'identifier — lui ou le groupe dont il fait partie — ou non. En effet, cet individu sait que ce renseignement lui a été demandé, qu'il est utilisé et qu'il est communiqué.

L'efficacité attribuée à l'anonymisation des données est chaudement disputée. Selon Cynthia Dwork, chercheuse et informaticienne réputée, « les données soumises à l'anonymisation ne deviennent pas anonymes ». Il n'est nullement certain qu'un renseignement dépersonnalisé ne pourra jamais redevenir personnel, bien au contraire, puisque ce risque est bien réel. Malgré tout cela, de nombreux organismes continuent de se tourner vers l'anonymisation lorsque vient le moment de communiquer leurs données.

Si un organisme choisit la stratégie de l'anonymisation, il est recommandé qu'il se penche sur les questions suivantes :

1. Prudence : ne pas surestimer l'efficacité de l'anonymisation;
2. Pratiques exemplaires et normes acceptées;
3. Approches normalisées;
4. Risques de préjudices;
5. Risque que les renseignements redeviennent personnels;
6. Application des techniques d'anonymisation;
7. Considérations ciblées en fonction du type de données.

1. **Prudence : ne pas surestimer l'efficacité de l'anonymisation** – Comme indiqué plus haut, le risque qu'un renseignement puisse redevenir personnel est bien réel. Les spécialistes de la sécurité et de la protection de la vie privée ont une piètre opinion des méthodes utilisées pour évaluer ce risque. Ils s'inquiètent également de l'insuffisance de la sécurité appliquée aux catégories de données, ainsi que du fait que la motivation, les connaissances et les ressources d'un adversaire voulant redonner leur caractère personnel aux données sont sous-estimées.

2. Pratiques exemplaires et normes acceptées – Il n'existe actuellement ni norme ni protocole d'application généralisée ou ayant valeur juridique en matière d'anonymisation des données. Aucun des cadres d'« utilisation fiable » assortie d'une responsabilité publique ayant été proposés jusqu'ici pour minimiser le risque qu'un renseignement puisse redevenir personnel n'a été adopté de façon générale. Certains domaines d'activité se sont dotés de normes et de pratiques exemplaires en matière d'anonymisation. À titre d'exemple, les *Pan-Canadian De-Identification Guidelines for Personal Health Information* publiées en 2007 par le docteur Khaled El Emam, défenseur avoué de l'anonymisation, sont souvent consultées au Canada lorsque vient le moment de communiquer à des tiers des renseignements médicaux.

3. Approches normalisées – Puisqu'il n'existe aucune approche normalisée en matière d'anonymisation, le commissariat ontarien à l'information et à la protection de la vie privée a publié en 2016 ses *De-Identification Guidelines for Structured Data*. Outre qu'elles énoncent une approche normalisée en matière d'anonymisation, ces lignes directrices calculent le risque que divers types de renseignements puissent redevenir personnels et présentent alors les techniques d'anonymisation appropriées. Loin de se vouloir la panacée pour les questions de protection de la vie privée soulevées par les données, ces lignes directrices proposent une approche systématique en matière d'anonymisation, à laquelle un recours répété permettra d'améliorer les pratiques de communication de données de l'organisme et de réduire les risques qu'elles font peser à la vie privée. D'autres exemples d'approche normalisée comprennent notamment le *Guide to Protecting the Confidentiality of Personally Identifiable Information* de l'American National Institute of Standards and Technology, et l'*Avis 05-2014 sur les techniques d'anonymisation* publié par le Groupe de travail sur l'Article 29 de l'Union européenne.

4. Risques de préjudices — Toute stratégie d'anonymisation devrait tenir compte des risques d'une communication accidentelle ou qu'un renseignement puisse redevenir personnel de quelque façon que ce soit. Ce dernier risque étant bien réel, l'organisme devrait

sérieusement s'interroger sur le bien-fondé de communiquer des renseignements sensibles sous forme dépersonnalisée, puisque ces derniers pourraient, s'ils redevenaient personnels, être d'une grande valeur à quiconque voudrait en tirer profit et, partant, porter atteinte à la vie privée d'une personne ou même lui nuire. Il faut se rappeler qu'il peut être difficile de distinguer un renseignement sensible (p. ex. de nature médicale ou financière) d'un autre qui ne l'est pas, puisque certains renseignements non sensibles peuvent le devenir une fois combinés à d'autres données.

5. Risque que les renseignements redeviennent personnels – Le niveau d'anonymisation nécessaire pour protéger la vie privée d'une personne devrait correspondre au risque que ses renseignements dépersonnalisés puissent redevenir personnels. L'évaluation de ce risque devrait faire appel à divers facteurs comme la nature sensible des renseignements, leur niveau de détail, le danger que pourrait courir la personne si ces renseignements étaient utilisés à mauvais escient, la valeur possible que les renseignements redevenus personnels auraient pour un tiers et la probabilité qu'un tiers s'efforce de rendre à nouveau personnels des renseignements dépersonnalisés.

6. **Application des techniques d'anonymisation** – Il existe plusieurs techniques d'anonymisation : le masquage d'identifiants, la modification de la taille de catégories équivalentes, le recours à des pseudonymes, le chiffrage de valeurs, l'ajout de « bruit » dans les données (comme l'insertion témoin), la suppression de colonnes de variables identificatrices, la généralisation de données sur plusieurs catégories, la protection différentielle de la vie privée et la suppression de quasi-identifiants. Il revient à l'organisme de choisir judicieusement les techniques les plus appropriées aux données visées.

7. Considérations ciblées en fonction du type de données – Les stratégies d'anonymisation varient selon le type de données visé :

a) **Données de localisation** – Il n'existe aucune preuve de l'efficacité de l'anonymisation de données de localisation. Voir à ce sujet notre *Explicatif : Données de localisation*.

b) **Mégadonnées** – Les mégadonnées renferment plusieurs champs distincts et les applications d'analyse en raffolent. Les mégadonnées sont donc devenues la norme, et leur anonymisation est rarement couronnée de succès puisqu'elles offrent trop de possibilités de croisement pour qui souhaite qu'elles redeviennent personnelles. Par conséquent et jusqu'à preuve du contraire, il faut présumer que l'anonymisation des mégadonnées est inefficace. Ce domaine étant en évolution, il faut en outre faire montre de prudence à l'égard des nouvelles approches ou normes qui pourraient être proposées.