Nouveau guide officiel ontarien sur l'anonymisation des renseignements personnels : un loupage épique

par Arnaud Palisson, PhD, CIPM

Le 30 octobre dernier, le *Commissaire à l'information et à la protection de la vie privée de l'Ontario* a publié un nouveau guide[1] en matière d'anonymisation des renseignements personnels. Il s'agit de la mise à jour d'un document de référence, dont la précédente mouture remonte à 2016.[2] Cette dernière s'avérait pertinente à bien des égards. Il me tardait donc de lire sa dernière version.

Comme en 2016, le nouveau guide se concentre sur l'anonymisation des **données structurées**.[3] C'est effectivement le champ de prédilection de l'anonymisation. En outre, pour simplifier et circonscrire le sujet, le document déclare se focaliser sur les données structurées sous forme de tableaux, avec lignes et colonnes ; tableaux dans lesquels une personne n'est présente qu'à une seule occasion.[4]

> Cette précision limite la portée du document. Mais c'était inévitable. En effet, le modèle du guide ontarien est basé sur le **k-anonymat**.[5] Lequel peut difficilement s'appliquer si des personnes sont présentes à plusieurs endroits dans une même base de données. Si le guide avait également traité de l'anonymisation basée sur des techniques de **confidentialité différentielle**, cette limitation n'aurait pas eu lieu d'être.[6] Mais il est vrai que cette matière est plus compliquée et qu'elle s'accorde mal avec la volonté pédagogique du guide de 2025.

---

[1] Information and Privacy Commissioner of Ontario, *De-identification Guidelines for Structured Data*, 2025 (ci-après abrégé "IPC Ontario"), https://www.ipc.on.ca/en/resources/de-identification-guidelines-structured-data

[2] Information and Privacy Commissioner of Ontario, *De-identification Guidelines for Structured Data*, 2016 https://www.ipc.on.ca/sites/default/files/legacy/2016/08/ Deidentification-Guidelines-for-Structured-Data.pdf

[3] Données stockées selon un format prédéfini, de façon à permettre leur traitement par un logiciel.

[4] *« The assumption is that there is only one row per individual*.»: IPC Ontario, *op. cit.*, p. 15.

[5] *Cf. infra*, § 3.2.2.

[6] Garfinkel, S. L. (2025). *Differential privacy.* The MIT Press. (Chap. 1, Section « Group Privacy »).

Fort d'un plus grand volume d'informations (il compte quatre fois plus de pages qu'en 2016), le nouveau document approfondit la doctrine du Commissaire de l'Ontario. Il énonce ainsi divers éléments frappés au coin du bon sens, explicitant certaines notions et façons de faire et renforçant ainsi leur bien-fondé. Plusieurs annexes du document, notamment les listes de vérification, constituent des outils intéressants.

Mais globalement, nous trouvons la version de 2025 décevante. En effet, à divers égards, le nouveau guide accumule approximations, faussetés voire égarements ; et ce, dans trois domaines fondamentaux :

1. les définitions de concepts-clés,
2. le processus d'anonymisation,
3. le calcul du risque de réidentification.

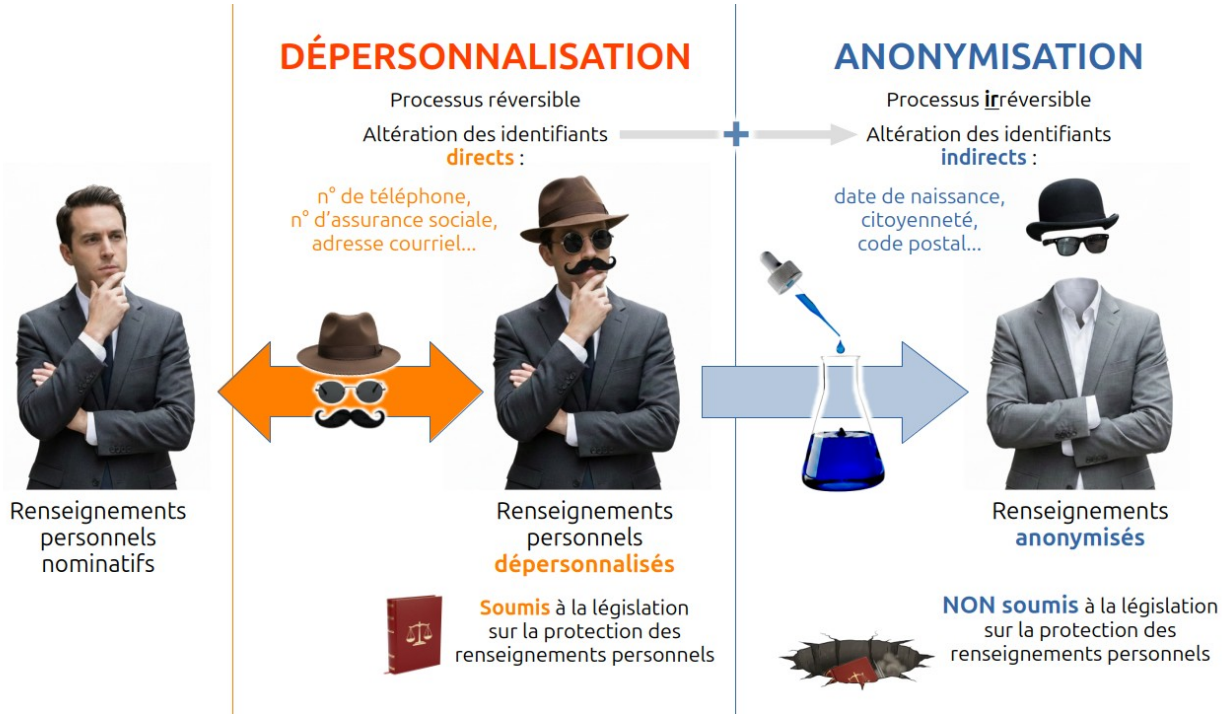
Cela nous paraît d'autant plus inquiétant que le nouveau guide du Commissaire ontarien est d'ores et déjà présenté dans l'industrie comme une référence canadienne sur ces trois éléments.[7] Aussi nous a-t-il semblé important de pointer du doigt ces problèmes conceptuels et méthodologiques.

1 Imprécision dans les définitions de concepts-clés

1.1 Rappel de la distinction entre anonymisation et dépersonnalisation

Avant de rentrer dans les imprécisions du guide, et afin de bien comprendre les problèmes posés par la terminologie qu'il utilise, il faut d'ores et déjà faire une distinction entre les trois notions essentielles que sont l'anonymisation, la dépersonnalisation et la désidentification.

[7] *Cf.* notamment : Canton, D. (2025, 23 octobre). *How de-identification protects personal information in Ontario*. Harrison Pensa. https://www.harrisonpensa.com/how-de-identification-protects-personal-information/ ;
Lacasse, A. (2025, 27 octobre). *Behind Ontario's IPC newly updated de-identification guidelines*. IAPP. https://iapp.org/news/a/ontario-s-ipc-updates-internationally-renowned-de-identification-guidelines ;
Michaluk, D., Vani, M., & Sternin, A. (2025, octobre). *De-identification of personal information and the new IPC Ontario guidelines*. Borden Ladner Gervais. https://www.blg.com/en/insights/2025/10/deidentification-of-personal-information-and-the-new-ipc-ontario-guidelines.



La désidentification *(schéma synthétique)*

1.1.1 Dépersonnalisation

La dépersonnalisation consiste en la suppression ou la transformation des identifiants directs d'une base de données – ou d'informations sous d'autres formes. Le but n'est pas d'empêcher drastiquement toute réidentification mais :

- de rendre difficile la réidentification à la plupart des personnes. Cela reste possible, mais il faut produire quelques efforts pour y parvenir.
- de permettre la réidentification à certaines personnes qui peuvent en avoir besoin, et ce, sans grand effort.[8]

Les renseignements dépersonnalisés sont encore des renseignements personnels, auxquels s'applique toujours la législation sur la protection de la vie privée.

1.1.2 Anonymisation

L'anonymisation est une modification d'une base de données de renseignements personnels par laquelle on supprime ou transforme les identifiants directs **et** indirects, de sorte qu'il soit très difficile de réidentifier des personnes dans la base de données une fois anonymisée. Dès lors, les

[8] Pour un exemple : *cf. infra*, § 1.3.

informations de la base de données anonymisées ne sont plus des renseignements personnels. La législation relative à la protection des renseignements personnels ne s'applique plus à elles.

L'anonymisation est souvent présentée dans les législations et les références de l'industrie comme un processus irréversible. Cela signifie pour le moins que ce processus **n'est pas prévu pour permettre une éventuelle réidentification**. Et que si réidentification il doit y avoir, ce sera avec difficulté et par des entités ayant une volonté spécifique et d'importantes capacités de réidentification.

Je ne m'appesantirai pas sur les définitions d'*identifiant direct* et d'*identifiant indirect*. Il suffit ici de retenir que :

- la dépersonnalisation est réversible pour certaines personnes ou entités ; son résultat contient encore des renseignements personnels et demeure donc soumis à la législation sur la protection des renseignements personnels.
- l'anonymisation est censée ne pas être réversible ; son résultat ne contient plus de renseignements personnels et n'est donc plus soumis à la législation ;

1.1.3 Désidentification

En français ,le terme *désidentification* recouvre **à la fois**[9] l'anonymisation et la dépersonnalisation. Certains auteurs considèrent d'ailleurs que le terme anglais *de-identification* est l'exact équivalent du français *désidentification*.[10]

---

[9] Commissariat à la protection de la vie privée du Canada. (2016, mai). *Document de discussion sur les améliorations possibles au consentement sous le régime de la Loi sur la protection des renseignements personnels et les documents électroniques*. https://www.priv.gc.ca/fr/mesures-et-decisions-prises-par-le-commissariat/recherche/consulter-les-travaux-de-recherche-sur-la-protection-de-la-vie-privee/2016/consent_201605/

Fasken. (2020, 21 septembre). *Aperçu techno-juridique des concepts de renseignements « dépersonnalisés » et « anonymisés » introduits dans le PL64*. https://www.fasken.com/fr/knowledge/loi-25/21-apercu-techno-juridique-renseignements-depersonnalises-anonymises

Martineau, J.T. *et al.* (2022, mars). *Enjeux éthiques de l'IA en santé*. Observatoire international sur les impacts sociétaux de l'IA et du numérique. https://www.obvia.ca/sites/obvia.ca/files/ressources/202203-OBV-Pub-Sant%C3%A9IA_4Ethique.pdf

[10] Hintze, M. (2016, 31 octobre). *GDPR through the de-identification lens*. Future of Privacy Forum. https://fpf.org/wp-content/uploads/2016/11/M-Hintze-GDPR-Through-the-De-Identification-Lens-31-Oct-2016-002.pdf

Sarah Kopper, S. Sautmann, A. Turitto, J. (2020, juillet). *Désidentification des données*. The Abdul Latif Jameel Poverty Action Lab. https://www.povertyactionlab.org/fr/resource/desidentification-des-donnees

Mais dans les diverses législations nord-américaines anglophones, *de-identification* est un terme passe-partout, aux contours flous. Selon les États (américains) et les provinces (canadiennes), *de-identification* peut correspondre à :

- une anonymisation,[11]
- une dépersonnalisation forte,[12] ou
- une simple dépersonnalisation.[13]

La législation de l'Ontario constitue un modèle de confusion en la matière.

1.2 *De-identification* et anonymisation

> On trouvera à la fin de cette section un schéma synthétique de la confusion terminologique en droit ontarien.

1.2.1 *De-identification* = anonymisation ?

Le guide ontarien définit ainsi la *de-identification* :

> *De-identification is the general term for the process of removing personal information from a record or dataset.*
>
> *De-identification protects the privacy of individuals because, once effectively deidentified, a dataset is considered to no longer contain personal information. (…)*
>
> *It is important to note that de-identification does not reduce the risk of re-identifying a dataset to zero. Rather, the deidentification process produces datasets for which the risk of re-identification is very low. To put it another way, de-identification is the removal of information that identifies an individual or could be used, either alone or with other information, to identify an individual based on what is reasonably foreseeable in the circumstances.*[14]

---

[11] *Personal Health Information Act* (Nouvelle-Écosse), art.3 (g), https://nslegislature.ca/sites/default/files/legc/statutes%20HTML/personal%20health%20 information.htm

[12] Code of Federal Regulations (États-Unis), § 164.514, https://www.ecfr.gov/ current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514

[13] *Government Bill (House of Commons) C-27 (44-1) - First Reading - Digital Charter Implementation Act*, 2022 – Parliament of Canada , article 2, https://www.parl.ca/documentviewer/en/44-1/bill/C-27/first-reading

[14] IPC Ontario, *op. cit.*, p. 2.

Le guide fait ensuite référence au terme *anonymization* mais… pour mieux écarter son utilisation. En effet, selon les auteurs, l'*anonymization* est une notion « très différente » de la *de-identification* :

> *In some jurisdictions, the term anonymization is also used as an alternative word for de-identification although we use these words to mean very different things in this guidance document. We will therefore avoid using the term anonymization, anonymized data, or anonymous data in this guide to minimize confusion.*[15]

Cette nette et péremptoire distinction a de quoi surprendre. Car la façon dont le guide utilise le terme *de-identification* recoupe intégralement la notion d'anonymisation telle qu'elle apparaît dans les législations au Québec[16], en Alberta[17] et dans les travaux parlementaires au fédéral.[18]

Certes, le guide ontarien rappelle qu'il est d'abord applicable dans le cadre de la législation ontarienne[19] :

> *This guidance document generally uses the term de-identification in a manner similar to Ontario's privacy laws and aligns with decisions, guidance and materials issued by the IPC.*

Mais cette précaution ne résout pas le paradoxe. Car, sur la définition de *de-identification*, le droit ontarien manque cruellement de clarté.

Ainsi, l'article 2 de la *Loi de 2004 sur la protection de renseignements personnels sur la santé* (dite PHIPA) utilise:

- le terme *de-identify* en version anglaise,[20]
- le terme *anonymiser* dans sa version française officielle.[21]

---

[15] IPC Ontario, *op. cit.*, p.6.

[16] *Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels*, ch. A-2.1, article 73, alinéa 2, https://www.legisquebec.gouv.qc.ca/fr/document/lc/a-2.1?langCont=fr#se:73

[17] *Protection of Privacy Act*, SA 2024, c P-28.5, article 1er , https://kings-printer.alberta.ca/1266.cfm?page=p28p5.cfm&leg_type=Acts&isbncln=9780779856954&display=html

[18] *Government Bill (House of Commons) C-27 (44-1)*, *op. cit.*, article 2 (1).

[19] IPC Ontario, *op. cit.*, p.5.

[20] *Personal Health Information Protection Act*, 2004, S.O. 2004, c. 3, Sched. A, section 2.

[21] *Loi de 2004 sur la protection des renseignements personnels sur la santé*, L.O. 2004, chap. 3, annexe A, article 2, https://www.ontario.ca/lois/loi/04p03#BK3

Donc, en Ontario, la lettre de la loi nous dit que *de-identification* (EN) et *anonymisation* (FR), c'est la même chose.

En outre, l'un des principaux auteurs du guide est l'universitaire spécialiste de l'anonymisation, Khaled El Emam. Actuellement en résidence universitaire au bureau de la Commissaire ontarienne, il utilise pour sa part le terme *anonymization* – notamment sur le site web dudit bureau.[22]

On comprend alors assez mal pourquoi le guide ontarien prétend que *de-identification* et *anonymization* sont deux notions « *very different* ».

1.2.2 *De-identification* = dépersonnalisation ?

Depuis une modification de la PHIPA en 2019, la *de-identification* est définie à l'article 2 d'une façon plus proche de l'anonymisation. Toutefois – et c'est là une seconde source de méprise – certaines dispositions de la PHIPA, notamment son article 11.2, rapprochent davantage la *de-identification* de la **dépersonnalisation**. En effet, cet article prohibe la réidentification de renseignements « anonymisés », tout en autorisant certaines entités à le faire.

Mais à quoi cela servirait-il d'autoriser expressément une entité à réaliser une réidentification de renseignements anonymisés, dès lors que :

1. le risque de réidentification est très faible ?
2. les entités prévues par la loi[23] ne sont pas censées disposer des capacités de mener une réidentification aussi difficile ?

Il nous semble donc, qu'aux termes de l'article 11.2, des renseignements « anonymisés » seraient plutôt des renseignements dépersonnalisés.

1.2.3 *De-identification* = anonymisation, avec liens entre les bases anonymisées !

Comme si ce n'était pas encore assez compliqué, l'ontarienne *Loi sur l'accès à l'information et la protection de la vie privée*[24] (dite FIPPA) a trouvé le moyen d'imaginer encore une autre approche de la *de-identification*. La définition du

---

[22] El Emam, K., *Embarking on my new journey as the IPC's Scholar-in-Residence*, 2 mai 2024, IPC.on.ca, https://www.ipc.on.ca/en/media-centre/blog/embarking-my-new-journey-ipcs-scholar-residence

[23] Notamment celles de l'article 11.2, alinéa (2), points 1. et 3.

[24] *Loi sur l'accès à l'information et la protection de la vie privée*, L.R.O. 1990, chap. F.31, https://www.ontario.ca/lois/loi/90f31

terme est similaire à celle de la PHIPA, le terme officiel en français est toujours *anonymisation*.[25]

Mais la loi crée le concept de *données anonymisées et liées* (c'est nous qui soulignons) :

> ***49.6*** *(1) Lorsqu'il recueille des renseignements personnels dans le cadre de la présente partie, le membre d'un service multisectoriel d'intégration des données ou d'un service ministériel d'intégration des données fait ce qui suit dès qu'il est raisonnablement possible de le faire dans les circonstances :*
>
> *1. Il crée un document renfermant la quantité minimale de renseignements personnels nécessaires afin d'anonymiser les renseignements et d'établir des liens entre ceux-ci et d'autres renseignements recueillis par le service.*
>
> *2. Il anonymise les renseignements personnels.*
>
> *3.* ***Si des liens doivent être établis, il établit des liens entre les renseignements personnels qui ont été anonymisés*** *en application de la disposition 2 et d'autres renseignements anonymisés au sein du service. (…)*

Comment la loi imagine-t-elle que l'on puisse établir des liens entre des renseignements personnels préalablement anonymisés, tout en assurant que ces renseignements ainsi recoupés **demeurent anonymisés** ? Cela contredit la nécessité, pour une anonymisation, d'atténuer les risques de corrélation et d'inférence.[26] Selon nous, cette disposition légale ne fait de sens qu'en contexte de dépersonnalisation.

1.2.4 *De-identification* = anonymisation !

Toutefois, selon la Commissaire ontarienne Patricia Kosseim, la *de-identification*, ce n'est pas de la dépersonnalisation mais bien de l'*anonymisation*[27] :

---

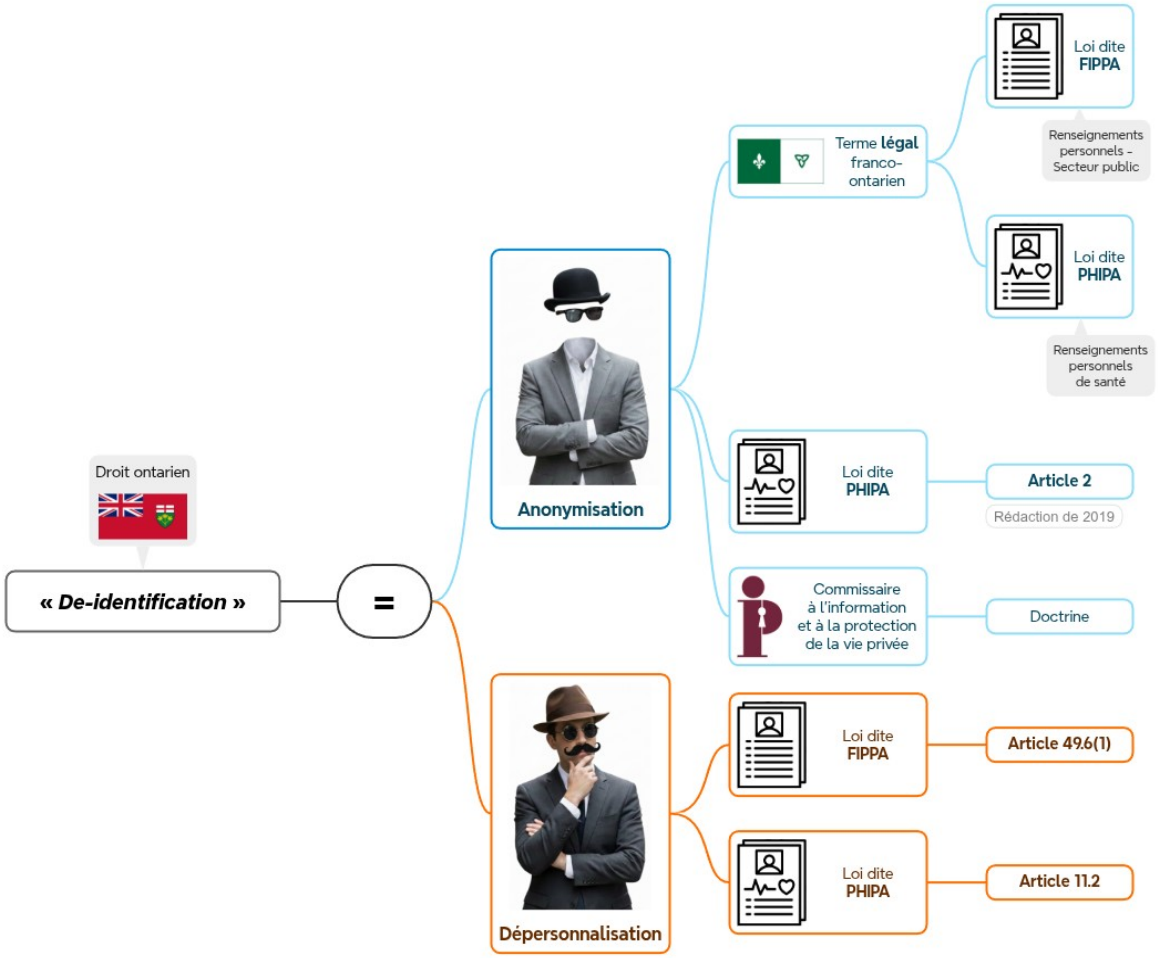
[25] Article 49.1 (2) de la *Loi sur l'accès à l'information et la protection de la vie privée.*

[26] *Cf. infra*, § 3.1.

[27] Kosseim, P. (2022, 17 mai). *Ripe for public debate: Legal and ethical issues around de-identified data.* Information and Privacy Commissioner of Ontario. https://www.ipc.on.ca/en/media-centre/blog/ripe-public-debate-legal-and-ethical-issues-around-de-identified-data ;
à comparer avec la version française du texte : https://www.ipc.on.ca/fr/centre-des-medias/blog/les-questions-juridiques-et-ethiques-entourant-les-donnees-anonymisees-un-debat-public-simpose

> *Ultimately, de-identification must be of the highest standard to ensure data cannot be re-identified and the privacy of individuals is protected. A robust de-identification governance process should include ongoing and regular re-identification risk assessments.*

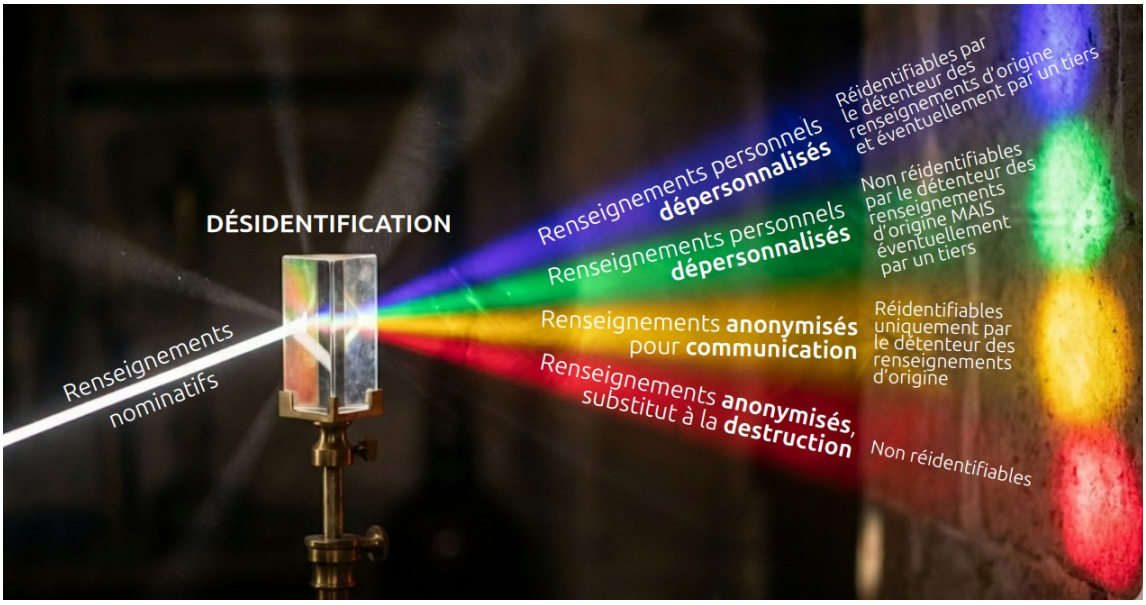
Bref, on l'aura compris, avec *de-identification*, les auteurs du guide utilisent un terme ambigu, pour définir une notion ambiguë de la loi ontarienne… Il aurait été beaucoup plus simple et univoque d'utiliser le terme *anonymization*, qui est d'ailleurs en train de s'imposer dans d'autres provinces canadiennes et au fédéral, ainsi que dans la recherche universitaire. Encore une occasion manquée.



Confusion en droit ontarien autour de la signification du terme de-identification

1.3 *Pseudonymization* et dépersonnalisation

Le document du Commissaire ontarien envisage la dépersonnalisation **uniquement comme une étape vers l'anonymisation**. Ce n'est pas faux, mais c'est extrêmement réducteur. En effet, au Québec et dans l'Union européenne, c'est un régime juridique complet de la dépersonnalisation qui a été créé, comblant une importante lacune. On dispose ainsi d'un spectre de l'altération des renseignements personnels dans la poursuite d'autres fins. Une façon de le représenter figure dans le schéma ci-dessous.



Quoi qu'il en soit, pour le guide ontarien, la dépersonnalisation au sens strict (simple étape intermédiaire vers l'anonymisation) est dénommée *pseudonymization* :

> ***De-identification*** *is the process of performing pseudonymization,* ***plus*** *transforming indirect identifiers that remain in the dataset following pseudonymization.*[28]

Le guide en donne les définitions suivantes :

> *Pseudonymization is the process of reducing the vulnerability of data by removing or transforming direct identifiers.*[29]
>
> *Pseudonymization is the process of transforming direct identifiers that exist within a dataset, such as names and residential addresses.*[30]

[28] IPC Ontario, *op. cit.*, p. 5.

[29] *Ibid.*, p.39.

[30] *Ibid.*, p. 5.

Or, le terme *pseudonymisation* existe également dans le *Règlement général sur la protection des données à caractère personnel* (RGPD) en Europe, défini à son article 4, 5) :

> *5) «pseudonymisation», le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.*[31]

Pseudonymisation renvoie en effet à la notion de pseudonyme : pour pseudonymiser,[32] on supprime des identifiants directs de la base de données, mais on y introduit pour chacune des personnes un *pseudonyme* afin de pouvoir les réidentifier aisément.

Prenons l'exemple d'une base de données médicales dépersonnalisées, transmise à un centre de recherche en oncologie, afin qu'il y calcule les probabilités d'occurrence d'un cancer.

Nom	Prénom	Sexe	Date de Naissance	Code postal	Glycémie (mmol/L)	Poids (kg)	Taille (cm)	Pathologie	...
Bouchard	Pierre	M	1960-08-14	G6V 5N4	5,5	88	175	Arthrose	
Fortin	Luc	M	1973-06-22	H4C 1V6	8,4	78	172	Diabète type 2	
Gagnon	Marie	F	1988-02-07	H3G 1M8	6,8	65	162	Pré-diabète	
Gauthier	Isabelle	F	1992-12-25	J0R 1K0	4,2	58	160	Anémie	
Leblanc	Sophie	F	1982-05-30	H1A 0A1	7,1	70	168	Diabète de type 2	
Morin	Mathieu	M	1980-08-04	H2W 1T7	6,0	92	185	Hypercholestérolémie	
Ouellet	Chantal	F	1999-10-10	J3Y 8G2	4,7	55	158	Allergies saisonnières	
Pelletier	Émilie	F	1987-01-17	G1S 1E4	5,8	63	165	Hypothyroïdie	
Roy	Sébastien	M	1995-12-11	J4K 2L9	4,9	75	182	Aucune	
Tremblay	Jean	M	1975-03-19	G1R 4Y8	5,2	82	178	Hypertension	

Base de données d'origine

[31] *Règlement général sur la protection des données à caractère personnel* (Union européenne), article 4, https://gdpr-text.com/fr/read/article-4/

[32] Office québécois de la langue française. (2002). *pseudonymiser*. Grand dictionnaire terminologique. Repéré à https://vitrinelinguistique.oqlf.gouv.qc.ca/fiche-gdt/fiche/8363289/pseudonymiser

Si les chercheurs du centre constatent que cette probabilité est très élevée chez l'un des patients dans la base, il est important d'être capable de l'identifier facilement afin de lui proposer rapidement de passer des tests de dépistage, voire de suivre un traitement adéquat. Mais comment faire, si on a supprimé les identifiants directs dans la base de données ? Aussi, **avant** de transmettre la base de données altérée, le détenteur des données originales crée une petite base de données annexe qui contient uniquement les renseignements identificatoires des personnes concernées ; on y ajoute un code de référence, contenant un identifiant spécial, propre à chaque individu. Il peut s'agir d'un numéro séquentiel, une chaîne de caractères aléatoires (*jeton*), un condensat cryptographique (*hash*), etc.

Nom	Prénom	Date de Naissance	Code de Référence
Bouchard	Pierre	1960-08-14	`9f2a8c4b7e1d`
Fortin	Luc	1973-06-22	`5d7b9e1a3c2f`
Gagnon	Marie	1988-02-07	`a1b2c3d4e5f6`
Gauthier	Isabelle	1992-12-25	`8f0e1d2c3b4a`
Leblanc	Sophie	1982-05-30	`c6d5b4a321e0`
Morin	Mathieu	1980-08-04	`2a4c6e80bdcf`
Ouellet	Chantal	1999-10-10	`13579bdfeca8`
Pelletier	Émilie	1987-01-17	`f0e1d2c3b4a5`
Roy	Sébastien	1995-12-11	`7a8b9c0d1e2f`
Tremblay	Jean-François	1975-03-19	`b5d3f1a2e4c6`

Table de correspondance entre identités et pseudonymes

Dans la base de données à dépersonnaliser, on supprime les renseignements identificatoires et on les remplace par cet identifiant spécial, qui tient lieu de *pseudonyme*.

Code de Référence	Sexe	Code postal	Glycémie (mmol/L)	Poids (kg)	Taille (cm)	Pathologie	…
`9f2a8c4b7e1d`	M	G6V 5N4	5,5	88	175	Arthrose	
`5d7b9e1a3c2f`	M	H4C 1V6	8,4	78	172	Diabète de type 2	
`a1b2c3d4e5f6`	F	H3G 1M8	6,8	65	162	Pré-diabète	
`8f0e1d2c3b4a`	F	J0R 1K0	4,2	58	160	Anémie	
`c6d5b4a321e0`	F	H1A 0A1	7,1	70	168	Diabète de type 2	
`2a4c6e80bdcf`	M	H2W 1T7	6,0	92	185	Hypercholestérolémie	
`13579bdfeca8`	F	J3Y 8G2	4,7	55	158	Allergies saisonnières	
`f0e1d2c3b4a5`	F	G1S 1E4	5,8	63	165	Hypothyroïdie	
`7a8b9c0d1e2f`	M	J4K 2L9	4,9	75	182	Aucune	
`b5d3f1a2e4c6`	M	G1R 4Y8	5,2	82	178	Hypertension	

Base de données dépersonnalisée par pseudonymisation

La table dépersonnalisée est alors transmise au tiers demandeur. Tandis que la table de correspondance est conservée sécuritairement par le détenteur de la base de données d'origine.

Par conséquent, **le terme *pseudonymisation* ne devrait être utilisé que pour cette seule forme de dépersonnalisation**. Or, ce n'est bien souvent pas le cas dans l'industrie de la protection des renseignements personnels, où les deux termes sont fréquemment et malencontreusement pris comme synonymes.

Le guide ontarien a pleinement conscience de la confusion.[33] Mais il maintient l'usage du terme spécifique *pseudonymization* comme vocable généraliste. Et dans son annexe F, consacrée aux techniques de *pseudonymization*, le document présente diverses techniques. Dont… la pseudonymisation ; et d'autres techniques qui ne relèvent pas de la pseudonymisation… On aurait aimé que, sur ce point, le guide ontarien ne rajoutât pas sa pierre au babélisme nord-américain. *So much for clarity.*

1.4 Quasi-identifiant

Autre controverse terminologique, le guide ontarien estime que le terme *quasi-identifiant* est un synonyme d'*identifiant indirect* :

> *Indirect identifiers (also sometimes called quasi-identifiers) are variables that can be used in combination with each other to identify an individual.*[34]

Nous ne sommes pas d'accord. En effet, tous les quasi-identifiants sont des identifiants indirects. Mais tous les identifiants indirects ne sont pas des quasi-identifiants :

- un identifiant indirect a un potentiel de réidentification ; une réidentification *serait possible* si et seulement si on lui adjoignait d'autres identifiants indirects pertinents;
- un quasi-identifiant permet une réidentification ; elle est *probable* car on dispose des autres identifiants indirects à cette fin.

Ainsi, on peut dire qu'un code postal est un identifiant indirect. Mais cela n'en fait pas ***ipso facto*** un quasi-identifiant.

---

[33] IPC Ontario, *op. cit.*, p. 80-81.

[34] *Ibid.*, p. 7.

Tandis que la date de naissance, le code postal domiciliaire et le sexe d'une personne **considérés ensemble** constituent des quasi-identifiants,[35] car ils permettent d'identifier des individus avec un fort degré de certitude.[36]

2 Un processus d'anonymisation "en tour d'ivoire"

Il existe deux grandes catégories de recours à l'anonymisation des renseignements personnels :

- **l'anonymisation-destruction** : elle intervient lorsque les renseignements personnels ne devraient plus être conservés par l'organisme détenteur car il n'en a plus besoin, ou parce que son calendrier de conservation lui impose de les détruire. Dans ce cas, on anonymise les renseignements personnels et on les conserve uniquement sous forme anonymisée. Les renseignements personnels initiaux sont alors détruits.
- **l'anonymisation-communication** : elle intervient lorsqu'une personne ou une organisation tierce demande qu'on lui transmette des renseignements personnels. S'il n'existe pas d'exception légale permettant ou exigeant la communication des renseignements personnels sous forme nominative ou dépersonnalisée, on se tourne vers l'anonymisation. Les renseignements anonymisés sont transmis au demandeur, mais les renseignements personnels d'origine sont conservés par leur détenteur dans leur forme initiale.

Le guide ontarien envisage **exclusivement la seconde** catégorie :

> *As the demand for data increases, data custodians require effective processes and techniques for removing personal information from the data so it can be used to draw important insights and improve services without compromising privacy or public trust. An important tool in this regard is deidentification.*[37]

---

[35] On peut même dire que le triplet constitue un quasi-identifiant.

[36] Sweeney, L. *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000. https://dataprivacylab.org/projects/identifiability/paper1.pdf.

[37] IPC Ontario, *op. cit.*, p. 1.

Il est frappant de constater que, **pour les auteurs de ce guide, l'anonymisation est d'abord et surtout l'affaire du spécialiste qui détermine les techniques d'anonymisation à appliquer**. Cela paraît sensé. Et pourtant, ça ne l'est pas.

Le guide ontarien prend soin de préciser que l'anonymisation est un processus de compromis entre le maintien de l'anonymat des personnes concernées et l'utilité résiduelle des renseignements anonymisés. Autrement dit, l'anonymisation est censée servir à quelque chose pour les destinataires des renseignements anonymisés. C'est bien pour cela qu'ils en font la demande.

Mais comment celui qui procède à l'anonymisation sait-il si les renseignements anonymisés seront véritablement utiles au destinataire ? On s'attendrait à ce que lesdits destinataires soient au moins consultés, en cours de processus, pour savoir si ce qu'on leur propose ressemble un tant soit peu à ce qu'ils souhaitent obtenir.

Eh bien, non. Le guide ontarien établit un standard : l'anonymisation est effectuée par le scientifique de données, depuis sa tour d'ivoire, en fonction de ce qu'il comprend de la demande quant aux **possibles utilisations ultérieures** des renseignements anonymisés.

C'est ce **spécialiste qui prend les décisions**, en se basant exclusivement sur son expérience et son expertise en matière d'anonymisation. Tout le chapitre 6 du guide, consacré au processus d'anonymisation, est organisé autour de cette idée que le scientifique de données effectuant l'anonymisation est le mieux placé pour savoir à quoi pourront servir les renseignements anonymisés.

En réalité, pour procéder à l'anonymisation, ledit scientifique de données en est rendu à jouer aux dés. Il sait d'expérience qu'ici, il va généraliser ; ici, il va masquer ; ici, il va perturber les identifiants. Il connaît la finalité globale prévue des renseignements anonymisés. Mais il n'a absolument aucune certitude que les données, une fois anonymisées, seront utiles à leur destinataire.

Cette conception du processus est intenable. Par exemple, qu'en est-il lorsque l'anonymisation est réalisée afin que les renseignements soient rendus publics (obligation de transparence, données ouvertes, etc.) ? Comment le scientifique de données peut-il connaître à l'avance les diverses utilisations qui seront faites de ces renseignements ?

Et que se passe-t-il si, en bout de ligne, le demandeur s'émeut de ne rien pouvoir tirer des renseignements anonymisés qui lui ont été remis ? On imagine que le scientifique de données lui répondra systématiquement, en substance :

> « Vous avez demandé des renseignements anonymisés ? Les voilà… Comment ça, ça ne fait pas votre affaire ?! Ah, mais, vous savez, l'anonymisation, c'est un compromis entre anonymat et utilité. Donc, si ça ne vous est pas utile, eh bien, ça arrive : c'est de l'anonymisation. Mais croyez-moi : j'ai fait de mon mieux pour que les données vous soient utiles. Je suis un expert, je sais ce qui est bon pour vous. Donc, si malgré cela, les données anonymisées ne vous conviennent pas, c'est que vous demandez l'impossible. »

Une démarche correcte d'anonymisation doit au contraire focaliser sur l'utilité résiduelle des renseignements anonymisés. Le demandeur sait que les données qu'il attend seront des renseignements anonymisés. Il a conscience qu'il y aura dans le processus une perte importante de précision dans les données. **Mais**, il faut lui offrir la possibilité d'exprimer le niveau de précision qu'il souhaite voir conservé pour les différents types de renseignements qui l'intéressent.

L'anonymisation est ensuite réalisée par un spécialiste. Lequel applique des techniques de dépersonnalisation puis d'anonymisation en cherchant à coller au plus près de la requête. Mais en ne dépassant pas le seuil de réidentification.[38]

On fournit alors au demandeur un échantillon de la base de données anonymisées (on s'assure que cet échantillon est également anonymisé).

- Si cet échantillon lui convient, l'intégralité de la base anonymisée lui est alors transmise.
- Si cela ne lui convient pas, on modifie la configuration de l'anonymisation et on génère une nouvelle fois la base anonymisée, en respectant le seuil de réidentification. Si cela ne suffit toujours pas, deux options se présentent à l'organisme détenteur des renseignements :
    - il explique au demandeur que l'on ne peut pas faire mieux, compte tenu du seuil de réidentification ; **OU**

[38] *Cf. infra*, § 3.5.

- il effectue une nouvelle anonymisation en collant davantage à la requête, mais en acceptant de dépasser le seuil de réidentification. Puis, il demande au responsable des risques de l'organisme s'il est possible de prendre un risque en transmettant la base de données au destinataire. Ici, le calcul de risque de réidentification est essentiel. Si le seuil de réidentification est, par exemple, établi à 10 %, selon que le risque de réidentification calculé sera de 12% ou de 78%, le responsable des risques sera plus ou moins enclin à accepter cette prise de risque.

Une telle approche de l'anonymisation a pour objectif de permettre au destinataire des renseignements anonymisés d'en faire bon usage, tout en respectant le seuil de réidentification.[39]

À l'opposé, le guide ontarien fait de l'anonymisation un traitement qui relève d'abord de la conformité légale, livrant l'utilité résiduelle à l'aléa technique des algorithmes. Si les renseignements anonymisés servent au demandeur, tant mieux ; sinon, tant pis.

Nous nous inquiétons que cette conception de l'anonymisation puisse devenir la référence canadienne. Elle ne rendra pas service aux organismes demandeurs des renseignements anonymisés.

Mais surtout, cette approche pénalisera les organismes détenteurs des renseignements personnels d'origine ! En effet, lorsque ces informations devront être détruites, lesdits organismes les auront mal faits anonymiser et ne seront pas en mesure, par la suite, de les utiliser comme ils l'avaient envisagé…

[39] C'est une approche similaire qui est recommandée pour l'anonymisation par confidentialité différentielle. *Cf.* Garfinkel, S.L., *op. cit.*, Chap. 1, section « Better Privacy Accounting Can Lower the Privacy Loss! ».

3 Le risque de réidentification

3.1 Les types de risque de réidentification

En matière d'anonymisation des renseignements personnels, il existe trois types de risque[40] :

- *identity disclosure* : l'analyse de la base de données anonymisées permet de réidentifier la personne ;
- *membership disclosure* : l'analyse permet d'établir si une personne déterminée figure ou non dans la base de données anonymisées ;
- *attribute disclosure* : l'analyse permet d'acquérir des informations sur une personne déterminée.

Cela recoupe le modèle traditionnel des trois risques à mitiger dans le cadre d'une anonymisation[41] :

- à l'*identity disclosure* correspond le risque d'individualisation (qui conduit à la réidentification) ;
- à la *membership disclosure* et à l'*attribute disclosure* correspondent, en s'y entremêlant, les risques de corrélation et d'inférence.

Il s'agit là d'une autre typologie des risques en matière d'anonymisation. Elle est tout-à-fait légitime et présente même un intérêt pédagogique indéniable.

Le guide prescrit à juste titre que l'atténuation du risque de réidentification passe par une atténuation de chacun de ces trois risques. Mais le document précise qu'il ne focalise que sur le premier d'entre eux, celui lié à l'*identity disclosure*.[42]

---

[40] Le guide ontarien ne parle pas de *risque* mais de *data vulnerability*. La terminologie s'avère ici doublement problématique. Tout d'abord, ces *vulnerabilities* ne sont pas des faiblesses de l'anonymisation, mais des résultats potentiels d'analyse de la base de données. On s'éloigne ici encore davantage des termes classiques de la gestion de risque. Par ailleurs, le guide réutilise plus loin le vocable de *data vulnerability* pour désigner une notion entièrement différente (*cf. infra*, § 3.2.2.). Tout cela ne va pas aider à la compréhension du modèle…

[41] *Cf.* notamment : Commission nationale Informatique et Liberté (CNIL), *L'anonymisation de données personnelles*, 19 mai 2020, https://www.cnil.fr/fr/technologies/lanonymisation-de-donnees-personnelles

[42] IPC Ontario, *op. cit.*, p. 12.

Ce n'est pas, loin de là, le seul document de référence à opter pour cette approche. Le guide ontarien s'en explique : les travaux sur les deux derniers risques sont moins bien développés que ceux consacrés au premier.[43] Eh bien, ç'aurait été l'occasion de commencer à inverser la tendance… La réputation du guide ontarien lui aurait permis de peser dans la balance. Au lieu de cela, il maintient le *statu quo* : les deux autres risques sont encore et toujours les parents pauvres de l'anonymisation des renseignements personnels.

3.2 Le calcul du risque de réidentification (*identity disclosure*)

3.2.1 Approche qualitative et approche quantitative

Une anonymisation a pour but de rendre impossible – du moins *très* difficile – une réidentification de personnes présentes dans une base de données. Afin de déterminer si une anonymisation est suffisamment efficace, on recourt à la notion de *risque de réidentification*.

Il est largement admis dans l'industrie, la littérature académique et certaines législations qu'une anonymisation digne de ce nom doit présenter un risque de réidentification **très faible**.

Certes, mais comment l'établir ? Il existe plusieurs façons de faire :

- certaines sont essentiellement **qualitatives** : on considère les différents facteurs de risque qui s'appliquent au cas d'espèce, et on les compare de façon empirique : certains facteurs sont plus élevés que d'autres, et on en tire une approximation sur l'ensemble des facteurs par rapport à un minimum absolu (tous les facteurs sont très faibles) et un maximum absolu (tous les facteurs sont élevés) ; et l'on détermine que, globalement, en l'espèce, le risque est *élevé*, *modéré*, *faible*, voire *très faible*. Cette démarche est évidemment très subjective ;
- d'autres méthodes sont essentiellement **quantitatives** : on essaie de transposer ces différents facteurs en chiffres, que l'on additionne (ou multiplie) ensuite entre eux. L'estimation est objective. Mais le problème se trouve en amont, dans la façon d'effectuer cette transposition chiffrée et d'intégrer les chiffres les uns aux autres. Et souvent, cette opération est réalisée de façon arbitraire, ce qui fausse largement l'estimation du risque.

[43] *Ibid.*, p. 16.

L'approche du guide ontarien relève de la seconde catégorie. Et ce n'est pas une réussite.

3.2.2 Une étrange équation du risque

Ainsi, selon les auteurs, le calcul du risque de réidentification obéit à l'équation suivante[44] :

$$Reidentification\ risk = data\ vulnerability \times probability\ of\ attack$$

Premier problème : si l'on choisit d'évaluer quantitativement le risque sur la base de ce qu'on appelle le *calcul du risque*, il convient traditionnellement de poser l'équation en ces termes :

$$Risque = Impact \times Probabilité$$

L'*impact* se définit comme les conséquences potentielles néfastes sur l'actif (information à protéger) qu'aurait l'événement redouté s'il devait survenir. Tandis que la *vulnérabilité* de l'information se définit comme la faiblesse du système traitant l'information ; faiblesse qui facilite la survenance de l'événement redouté. *Impact* et *vulnérabilité* sont donc des facteurs de risque totalement distincts. Et ils ne sont pas interchangeables.

En outre, la probabilité d'une attaque sur des données dépend de la vulnérabilité du système qui traite ces données. Autrement dit, la vulnérabilité est comprise dans la probabilité. Dès lors, vouloir **calculer le risque en multipliant la vulnérabilité par la probabilité ne fait aucun sens**.

Mais le guide va plus loin : il donne de *data vulnerability* une définition particulière :

> *Conceptually, you can think of the vulnerability as a probability of re-identification that can be assigned to each record in a dataset.*[45]

Pour comprendre ce que signifie cette phrase, il est nécessaire d'expliquer ici que le guide ontarien fonde sa notion de *data vulnerability* sur ce que l'on appelle le ***k*-anonymat**. Cela signifie que l'on évalue le niveau d'anonymisation d'une base de données selon le nombre de personnes qui, dans la base anonymisée, partagent exactement les mêmes

[44] IPC Ontario, *op. cit.*, p. 9.

[45] *Ibid.*, p. 31.

caractéristiques. Il est possible que plusieurs groupes de personnes partagent les mêmes caractéristiques ; mais pas les mêmes que dans les autres groupes. On va alors considérer le groupe qui compte le moins de personnes. Ce nombre de référence est la variable ***k*** – d'où le terme de ***k*-anonymat**.

La probabilité minimale de réidentification de personnes dans la base anonymisée s'écrit :

$$P_{Réid°} = \frac{1}{k}$$

Dès lors, l'équation de la page 9 du guide pourrait s'écrire :

reidentification risk = probability of re-identification × probability of attack

soit :

$$R_{Réid°} = P_{Réid°} \times P_{Attaque}$$

Le risque ne serait donc plus qu'une question de <u>probabilité</u>. *Exit* l'impact. *Exit* la vulnérabilité. Quelle étrange conception du risque !

3.3 Probabilité d'attaque de réidentification sur une base de données anonymisées publique

Le guide énonce plus loin les cinq principaux cas d'usage de l'anonymisation des renseignements personnels. L'un d'entre eux est la <u>publication en données ouvertes</u>.

Dans de tels cas, le document dispose :

> *Data vulnerability evaluation for public release errs on the conservative side.*[46]
>
> *If the data is to be made public then you always assume that an attack is certain to occur (i.e., the likelihood of an attack=1).*[47]

Et même si l'organisme divulgateur met fin à leur publication :

> *Organizations should generally assume that a re-identification attack is certain to occur either immediately or at any future time, as parties will generally be able to maintain local copies of released data.*[48]

[46] *Ibid.*, p. 40.

[47] *Ibid.*, p. 9.

[48] *Ibid.*, p. 19.

Il s'agit là, selon nous, d'une **sérieuse erreur conceptuelle**. Ce n'est pas PARCE QU'une base de données anonymisées est rendue publique qu'elle fera *ipso facto* et nécessairement l'objet d'une attaque en réidentification ! La probabilité ici, c'est l'éventualité que la base anonymisée soit attaquée si elle est rendue publique. C'est très différent. Car cette éventualité varie selon la base de données considérée.

En effet, selon vous, laquelle de ces deux bases de données anonymisées rendues publiques a la plus grande probabilité de subir une attaque de réidentification ?

- *Statut vaccinal des ouvriers œuvrant à la cueillette des olives en Basse-Provence entre 1956 et 1960*
- *Incidence de cancer en phase terminale chez les chefs d'État et chefs de gouvernement des 186 États reconnus par l'ONU, entre 2020 et 2025*

Quoi qu'il en soit, pour les auteurs du guide ontarien, en situation de base de données anonymisées à diffusion publique, la formule de calcul du risque est ainsi libellée[49] :

$$\textit{For public releases, risk of re-identification} = \textit{data vulnerability} \times 1$$

Autrement dit, pour les bases de données publiques :

$$\textit{Risque} = \textit{Vulnérabilité}$$

Cette conception est en **complète opposition avec le *Règlement québécois sur l'anonymisation des renseignements personnels***. Lequel prescrit[50] :

> *(…) les résultats de l'analyse doivent démontrer, en tenant compte notamment des éléments suivants, que les risques résiduels de réidentification sont très faibles:*
>
> *1° les circonstances liées à l'anonymisation des renseignements personnels, notamment les fins pour lesquelles elle entend utiliser les renseignements anonymisés;*
>
> *2° la nature des renseignements;*
>
> *3° le critère d'individualisation, le critère de corrélation et le critère d'inférence;*

[49] *Ibid.*, p. 40.

[50] A-2.1, r. 0.1 – *Règlement sur l'anonymisation des renseignements personnels*, article 7, alinéa 3, https://www.legisquebec.gouv.qc.ca/fr/document/rc/A-2.1,%20r.%200.1%20/

> *4° les risques que d'autres renseignements raisonnablement disponibles, notamment dans l'espace public, soient utilisés pour identifier directement ou indirectement une personne;*
>
> *5° les moyens nécessaires pour réidentifier les personnes, notamment en considérant les efforts, les ressources et le savoir-faire requis pour mettre en œuvre ces moyens.*

Ces dispositions réglementaires s'appliquent indistinctement aux diffusions non publiques et publiques de bases de données anonymisées. C'est cette conception du risque de réidentification que nous prônons.

3.4 Calcul de la probabilité de réidentification dans les bases de données **non** publiques

Pour les bases de données non publiques, la proposition de calcul du risque de réidentification se révèle moins indigente de prime abord, car elle intègre la probabilité d'une attaque en réidentification.

Selon les auteurs :

> *There are three types of attacks that can be modeled and evaluated: deliberate insider attack, inadvertent recognition of an individual in the dataset by an acquaintance, and data breach.*
>
> *The evaluation is the probability that the attack will occur, and that probability can depend on the context and controls.*

Pour expliquer sa méthode, le guide propose l'annexe E, dans laquelle les trois risques de réidentification correspondants sont savamment disséqués. Malheureusement, dans les trois cas, la méthode de calcul pèche par son manque de pertinence et surtout… de logique.

3.4.1 Attaque par une menace intérieure

Dans le premier type d'attaque, la réidentification est menée par un individu travaillant pour l'organisme détenteur de la base anonymisée – non publique.

Pour les auteurs du guide, la probabilité que cet individu lance une attaque de réidentification dépend de deux éléments :

1. le niveau de protection de la confidentialité de la base de données anonymisées,
2. la volonté et les capacités de l'individu à mener l'attaque contre cette base.

Le guide fournit ensuite le tableau ci-contre. Il explique que, par exemple :

- si les mesures de protection sont modérées, et
- si la menace est modérément capable et motivée,

alors la probabilité de **survenance** de l'attaque est de **0,3**.

Autrement dit, dans de telles conditions, **chaque base de données anonymisées produite par l'organisme a environ une chance sur 3 d'être la cible d'une attaque en réidentification**.

Privacy, Security, and Contractual Controls	Motives and Capacity	Probability of Re-Identification Attack
High	Low	0.15
	Medium	0.2
	High	0.25
Medium	Low	0.25
	Medium	0.3
	High	0.4
Low	Low	0.4
	Medium	0.5
	High	0.6

Figure 26: Assessment of controls and motives and capacity.

Encore pire : si l'individu est peu motivé/peu capable et que les mesures de sécurité sont faibles, chaque base de données anonymisée a… 40% de chances de subir une attaque en réidentification !

C'est **absurde**. Tout d'abord, il s'agit d'une déclinaison de l'idée vue plus haut qu'une base de données publique a 100% de chances de subir une attaque de réidentification. Elle n'est pas plus fondée quand elle est, comme ici, transposée aux bases de données non publiques.

Ensuite, le guide ontarien confond ici deux notions :

- la probabilité de **survenance** d'une attaque – qui dépend largement :
  - du contenu de chaque base de données et
  - de la volonté de l'attaquant ;
- la **réussite** de l'attaque, qui dépendra :
  - des mesures de protection de la base de données et
  - des capacités de l'attaquant.

Cette confusion invalide totalement la méthode de calcul proposée.

3.4.2 Risque de réidentification d'une personne par inadvertance par un individu connaissant la personne concernée

La seconde possibilité d'attaque de réidentification sur une base de données non publique nous est présentée comme suit :

> *In addition to deliberately attempting a reidentification attack, the recipient of a nonpublic data release may also inadvertently re-identify one or more individuals. This could happen if, while analyzing the data, they recognize a friend, colleague, family member or acquaintance. The probability of such an attack occurring is equal to the probability of a random recipient knowing someone in the dataset.*[51]

Ici, les auteurs font montre d'une expertise dans le calcul des probabilités. Il nous est même expliqué le fameux *chiffre de Dunbar*, intégré dans une équation savante. Mais leur hypothèse de travail pose deux problèmes.

Tout d'abord, la probabilité que le réidentificateur connaisse quelqu'un dans la base est calculée SANS prendre en compte la taille de la base de données. Nous trouvons cela surprenant. En effet, si la base compte 2 personnes, la probabilité que le réidentificateur connaisse quelqu'un qui y figure est NETTEMENT plus faible que si la base contient 50 millions de personnes.

Ensuite, quel est l'enjeu que l'employé reconnaisse la personne ? Pour comprendre cet autre paradoxe, prenons l'exemple d'un chercheur dans un institut de recherche sur le cancer, au sein d'un grand centre hospitalier du Québec. Un jour, il reçoit une base anonymisée de tous les patients atteints de cancer du cerveau dans la province.

Or, un de ses amis est atteint d'une telle pathologie. Et là, de 2 choses l'une :

- Soit le chercheur **sait** que son ami est atteint d'un cancer au cerveau :
    - Il n'y a alors pas de reconnaissance par inadvertance. La réidentification de la base ne lui apprendra pas ce qu'il sait déjà. Mais elle pourrait lui apprendre d'autres informations sur son ami (on est alors en *attribute disclosure*, et non en *identity disclosure*).[52]

[51] *Ibid.*, p. 75.

[52] *Cf. supra*, § 3.1.

- Le chercheur peut ainsi être plus motivé à procéder à une attaque :
  - la probabilité d'attaque est donc nettement supérieure ;
  - mais si le chercheur réussit sa réidentification, celle-ci n'aura pas été obtenue par inadvertance.

- Soit le chercheur **ne sait pas** que son ami est atteint d'un cancer au cerveau. Et là, se présentent deux hypothèses exclusives l'une de l'autre.
  - Soit le chercheur reconnaît son ami dans la base anonymisée.
    Mais si 15 ou 20 personnes partagent les mêmes caractéristiques que son ami et qu'il ignore que ledit ami est atteint d'un cancer au cerveau, comment le chercheur peut-il le reconnaître ?
  - Soit il reconnaît son ami, par inadvertance, dans la base **APRÈS** avoir effectué la réidentification. Dans ce cas, le fait qu'il connaisse la personne n'entre pas en jeu dans le calcul du risque de réidentification.

Ce second type d'attaque proposé par le guide ontarien se revèle donc inconséquent.

3.4.3 Fuite de données

La dernière hypothèse avancée dans le guide est celle d'une *data breach*, autrement dit un incident de sécurité au cours duquel une base de données anonymisées est consultée, volée, divulguée ou utilisée par une personne ou une entité non autorisée.

Pour les auteurs :

> *the probability of such an attack occurring is equal to the probability of a breach occurring at the recipient's facilities.*[53]

Autrement dit, selon eux, toute personne ayant accès à une base de données anonymisées va **nécessairement** tenter une attaque de réidentification.

[53] IPC Ontario, *op. cit.*, p.76.

En gestion des risques, une telle position n'est pas tenable :

- si l'incident est intentionnel, la probabilité d'une attaque de réidentification n'est pas négligeable (tout dépend du contenu allégué de la base), mais elle n'est certainement pas *très élevée*. En effet, rien ne prouve de prime abord que celui qui accède à la base de données anonymisées aura nécessairement la volonté de mener à bien une réidentification – même s'il en a effectivement les capacités ;
- si l'incident survient par inadvertance, la probabilité est *très faible* que l'entité qui accède aux données aura, ET la volonté, ET les capacités de mener à bien une réidentification.

À diverses reprises, les auteurs du guide précisent qu'ils ont une approche conservatrice du risque de réidentification. Comprenez : « Ne prenons pas de risque avec le calcul du risque. » Selon nous, cette approche conservatrice est inappropriée. De deux choses l'une :

- soit on se situe dans une gestion des risques, et on estime le risque de façon un tant soit peu rationnelle ;
- soit on imagine systématiquement le *worst case scenario*. Auquel cas, ce n'est plus une approche de risque, mais un simple exercice de conformité.

3.5 Seuil de réidentification et risque de réidentification

Le guide propose également une méthode de calcul du *seuil de réidentification*. Si le risque calculé est inférieur à ce seuil, on considérera que la base de données est correctement anonymisée.

Pour le calcul de cette valeur de référence, le guide recourt à la notion de *privacy invasion*[54], qui regroupe quatre facteurs :

> *The level of privacy invasion considers the potential privacy-related harms to individuals resulting from the release of the data, including potential injury or harm due to inappropriate processing. The level of invasion of privacy is a function of different factors, including:*

[54] Le vocable *privacy invasion* n'est pas très heureux. En effet, le 3e facteur avancé (nombre de personnes concernées par la réidentification) n'est pas un indicateur du niveau d'intrusion dans la vie privée. Lequel niveau d'intrusion s'évalue au niveau individuel et non collectif.

- *the sensitivity of the information*
- *the scope and/or level of detail of the information*
- *the number of individuals that would be affected by a successful re-identification attack*
- *the potential harms or injuries to individuals in the event of a breach or inappropriate use*

The result of the invasion of privacy assessment should be a qualitative value typically in the range of low, medium or high. However, the amount of deidentification that you need to apply to a dataset is quantified numerically. To bridge this divide, once you have assessed the invasion of privacy value, you now have to translate the result into a numerical value, representing the amount of deidentification required proportionate to that level of risk. This identifiability threshold represents, in general, the minimum amount of de-identification that must be applied to a dataset for the level of reidentification risk to be considered very low. At that point and beyond, the data will be considered de-identified and no longer personal information. Accordingly, it forms the baseline against which to compare your calculations concerning deidentification going forward.[55]

Pour les auteurs du guide, le seuil de réidentification est calculé en fonction du niveau d'intrusion dans la vie privée que présentent les renseignements personnels dans **chaque** base de données. Le seuil de réidentification est donc une variable.

Le document précise :

You may use the table in figure 11 as a guideline in determining what is considered a very low risk of re-identification for different datasets, depending on their invasion of privacy values.

Invasion of Privacy Values	Re-identification Risk Threshold (very low)	Cell Size Equivalent
Low	0.09	11
Medium	0.075	15
High	0.05	20

Figure 11: Invasion of privacy assessment to determine the (very low) risk threshold.

[55] IPC Ontario, *op. cit.*, p. 39.

> *The values listed in the table are consistent with data release precedents across Canada and internationally. The table also includes the cell size or group size equivalent for each reidentification risk threshold. For example, a dataset with a re-identification risk threshold of 0.09 means that there must be no fewer than 11 rows (1 divided by 0.09) in which the indirect identifiers have the same values (i.e., there must be no fewer than 11 individuals who share the same characteristics or traits before the risk of reidentifying any one of them surpasses the risk threshold).*[56]

Autrement dit, si, dans la base de données originale, les facteurs d'intrusion dans la vie privée sont globalement à un niveau élevé, il faut anonymiser la base de données de sorte qu'il y ait toujours au moins 20 personnes qui partagent exactement les mêmes caractéristiques dans la base de données anonymisée ($k_{Seuil} = 20$).

Si les facteurs d'intrusion sont globalement faibles, le seuil de réidentification passe à au moins 11 personnes partageant les mêmes caractéristiques ($k_{Seuil} = 11$).

Une fois le seuil établi, on va calculer le risque de réidentification de la base telle qu'on l'a anonymisée. Ce calcul va nous donner la valeur de k de la base telle qu'elle est anonymisée (k_{Base}).

Pour une même base de données :

- si $k_{Seuil} = 11$ et $k_{Base} = 28$, tout va bien : dans la base anonymisée, il y a nettement plus de 11 personnes (28) qui partagent les mêmes caractéristiques ;
- si $k_{Seuil} = 11$ et $k_{Base} = 7$, alors la base de données est mal anonymisée. Il faut transformer davantage les identifiants, puis recalculer le risque jusqu'à ce que $k_{Base} \geqslant 11$.

On comprend l'idée. Mais il y a un (gros) problème :

- k_{Seuil} est calculé en se basant sur les facteurs d'intrusion dans la vie privée. On considère ici l'**impact** potentiel sur la vie privée APRÈS une éventuelle réidentification ;

[56] *Ibid.*, p 40.

- k_{Base} est calculé SANS se baser sur ces facteurs, en considérant uniquement le degré d'anonymisation de la base. On considère ici la **vulnérabilité** de la base AVANT une éventuelle réidentification.

Bref, avec sa méthode de calcul du seuil de réidentification, le guide ontarien nous invite à comparer des pommes avec des oranges…

*
* *

La nouvelle version du guide *De-identification Guidelines for Structured Data* constitue une importante refonte de la doctrine du Commissaire à l'information et à la protection de la vie privée de l'Ontario concernant l'anonymisation des renseignements personnels.

Certains concepts y sont davantage développés, mieux explicités et théoriquement renforcés par rapport à l'édition de 2016. En revanche, on regrettera amèrement que certains aspects essentiels de la discipline soient présentés, selon nous, de façon largement erronée.

1. On relève en effet des approximations dans les définitions de plusieurs termes cruciaux.
2. De plus, le processus d'anonymisation présenté se focalise sur la personne du scientifique de données chargé d'effectuer une anonymisation comme bon lui semble, ne s'intéressant que marginalement à l'utilité résiduelle que peut en attendre le demandeur de la base de données anonymisées.
3. Enfin, le modèle quantitatif proposé pour le calcul du seuil et du risque de réidentification nous paraît largement inadapté – c'est un euphémisme.

Il en ressort que le guide ontarien tend à faire de l'anonymisation un processus ancré dans la recherche d'une confortable conformité légale. En ce sens, il s'oppose diamétralement à l'esprit de certaines législations – hors Ontario. Il est donc, à nos yeux, d'autant plus regrettable que le landerneau canadien de la protection des renseignements personnels présente d'ores et déjà le guide ontarien de 2025 comme une référence pour les organismes désireux de mettre en place un processus d'anonymisation. On peut (et on doit) mieux faire. ■