UNIVERSITÉ DE LAUSANNE

ECOLE DE DROIT
FACULTÉ DE DROIT, DES SCIENCES CRIMINELLES
ET D'ADMINISTRATION PUBLIQUE

# Distinction entre données personnelles et anonymes : *cadre juridique, réponses techniques et conséquences sur le partage*

MÉMOIRE

présenté

par

**Jasmine Brantschen**

sous la direction de

**Prof. David-Olivier Jaquet-Chiffelle**

Lausanne, 3 juin 2024

# Table des matières

*Tous mes remerciements vont au Professeur David-Olivier Jaquet-Chiffelle pour avoir accepté de diriger le présent mémoire ainsi que pour la confiance accordée dans le choix du sujet. Je tiens également à remercier son assistante, Delphine Sarrasin, qui a accepté de superviser la partie juridique de ce travail pluridisciplinaire. Je remercie également Mikhaël Salamin, Conseiller à la protection des données de l'Université de Lausanne, pour son encadrement et les nombreux échanges très intéressants que nous avons eus sur le sujet. Finalement je tiens à exprimer ma gratitude à ma chère amie, Lirjona Dermaku, pour la relecture attentive !*

# Bibliographie


*Agencia Española de Protección de Datos* (AEPD), *10 misunderstandings related to anonymization,* avril 2021, disponible sous: https://www.edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf
(cite : AEPD *10 Misunderstandings)*

BARTH-JONES Daniel, *The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now,* Juillet 2012, disponible sous: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2076397

BENHAMOU Yaniv/COTTIER Bertil, *Loi sur la protection des données, petit commentaire,* Bâle 2023.

DWORK Cynthia/ MCSHERRY Frank/ NISSIM Kobbi/ SMITH Adam, *Calibrating Noise to Sensitivity in Private Data Analysis,* V 3876 Theory of Cryptography, 2006.

DWORK Cynthia/ ROTH Aaron, *The Algorithmic Foundations of Differential Privacy,* Foundations and Trends in Theoretical Computer Science, vol. 9, nos. 3–4, pp. 211–407, 2014.

ERARD Frédéric, *Les données codées dans le context de la recherche : personnelles ou anonymes ?,* AJP/PJA mai 2021.
(cité : ERARD (code))

ERARD Frédéric/ HEUSGHEM Mathilde/ PARISATO Clément, *Recherche biomédicale et Open Data, perspective en droit suisse,* jusletter du 30 janvier 2023, disponible sous : https://jusletter.weblaw.ch/fr/dam/publicationsystem_leges/erard-et-al_jl-30-jan-2023/recherche-biomedical_4b9e008242/Jusletter_recherche-biomedical_4b9e008242_fr.pdf

*European Union Agency for Cybersecurity* (ENISA), *Pseudonymisation techniques and best practices – Recommandations and shaping technology according to data protection and privacy provisions*, novembre 2019.
(cite: ENISA *pseudonymisation)*

FERRARI HOFER Lorenza/ GEORG PICHT Peter/ MATHYS Roland/ MAMANE David, *Données et bases de données: Cadre juridique et pertinence pour le marché,* in Jusletter octobre 2021, disponible sous :
https://www.swlegal.com/media/filer_public/be/aa/beaa996c-f8ba-499f-8389-2e55e4fc3933/nl_october_2021_data_french.pdf
(cité : FERRARI HOFER/ GEORG PICHT/ MATHYS/ MAMANE)

FORS Guide, *Quantitative data anonymization: practical guidance for anonymizing social science data,* No, 23, Version 1.0, Mars 2024, disponible sous:
https://serval.unil.ch/resource/serval:BIB_13CC50576B31.P001/REF.pdf

ERARD Fréderic, *La protection des données dans la recherche*, in Sylvain Métille (éd.), Stämpfli Editions SA, Berne 2024, p. 1ss.

Groupe de travail « Article 29 » sur la protection des données, *Avis 4/2007 sur le concept de données à caractère personnel,* adopté le 20 juin, disponible sous : https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_fr.pdf
(cité : WP 136)

Groupe de travail « Article 29 » sur la protection des données, *Avis 5/2014 sur les Techniques d'anonymisation,* adopté le 10 juin 2014, disponible sous : https://www.cnil.fr/sites/cnil/files/atoms/files/wp216_fr.pdf
(cité : WP 216).

Groupe de travail « Article 29 » sur la protection des données, *Opinion 06/2013 on open data and public sector information (« PSI ») reuse,* adopté le 5 juin 2013, disponible sous : https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp207_en.pdf
(cité WP 207).

Groupe de travail « Article 29 » sur la protection des données, *Opinion 03/2013 on purpose limitation,* adopté le 2 avril 2013, disponible sous : https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf
(cité : WP 203).

Information and Privacy Commissioner of Ontario (IPC), *De-identification Guidelines for Structured Data*, June 2016, disponible sous : https://www.ipc.on.ca/sites/default/files/legacy/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

Information Commissioner's Office (ICO), *Anonymisation : managing data protection risk code of practice,* novembre 2012, disponible sous: https://ico.org.uk/media/1061/anonymisation-code.pdf
(cité : ICO *Code of practice)*

ISO/IEC 20889:2018: Privacy enhancing data de-identification terminology and classification of techniques.

JACOT-GUILLARMOD Emilie, *US Program: le transfert de données clients pseudonymisées,* in Jusletter 5 octobre 2018, disponible sous : https://www.lawinside.ch/660/#:~:text=Les%20donn%C3%A9es%20pseudonymis%C3%A9es%20de%20fa%C3%A7on,identification%20est%20effectivement%20rendue%20impossible.

JACOT-GUILLARMOD Emilie/HIRSCH Célian, *Les données bancaires pseudonymisées – Du secret bancaire à la protection des données,* SZW-RSDA 2/2020, p. 151ss.

JOTTERAND Alexandre, *Des données personnelles pseudonymisées transférées à un tiers deviennent-elles anonymes ?,* Justletter du 13.06.2023, disponible sous : https://www.swissprivacy.law/?pdf=6146
(cité : JOTTERAND (arrêt TUE))

JOTTERAND ALEXANDRE, *Personal Data or Anonymous Data: where to draw the line (and why)?,* in Jusletter du 15 août 2022, disponible sous: https://jusletter.weblaw.ch/fr/juslissues/2022/1119/personal-data-or-ano_173939252d.html__ONCE&login=false

KAMATH Gautam, *Lecture 1 – Some Attempts at Data Privacy*, 2020, disponible sous: http://www.gautamkamath.com/CS860notes/lec1.pdf

KNIOLA Lukasz, *Plausible Adversaries in Re-Identification Risk Assessment*, 2017, disponible sous: https://www.lexjansen.com/phuse/2017/dh/DH09.pdf

LAKOMAA Erik/ KALLBERG Jan, *Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs,* août 2013.

MARMIER Auriane/METTLER Tobias, *Proposition pour la publication des données ouvertes publiques : working paper de l'IDHEAP,* 2019, disponible sous : https://serval.unil.ch/resource/serval:BIB_6E7703DA604E.P001/REF.pdf

MEIER Philippe, Commentaire d'arrêt JdT 2011 II p. 446, disponible sous : https://www.swisslex.ch/fr/doc/clawrev/aba5961e-78ac-470a-ab60-293078ff57a1/search/201247983

MEIER Philippe, *Le défi de Big Data dans les relations entre privés. Avec quelques réflexions de lege feranda,* dans Big Data et droit de la protection des données, 2016.
( cité : MEIER (Big Data) )

MEIER Philippe, Protection des données, *Fondements, principes généraux et droit privé,* Berne 2011.

MEIER Philippe/MÉTILLE Sylvain (édits), *Loi fédérale sur la protection des données :* commentaire romand, Bâle 2023.

NARAYANAN Arvind/ SHMATIKOV Vitaly, *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset,* 5 février 2008, disponible sous: https://arxiv.org/pdf/cs/0610105

Office fédéral de la santé publique (OFSP), *Questions fréquentes sur les activités de recherche, l'anonymisation et la transmission de données*- dans le cadre de l'exécution de la loi fédérale sur l'enregistrement des maladies oncologies (LEMO), 20 octobre 2020, disponible sous : https://m4.ti.ch/fileadmin/DSS/DSP/RCT/pdf/pdf_privati/FAQ_Datenweitergabe_Anonymisierung_Forschung.pdf
(cité : OFSP Questionnaire)

Office fédéral de la statistique (OFS), *Conditions d'utilisation du cadre d'échantillonnage,* 08.06.2022, disponible sous : https://dam-api.bfs.admin.ch/hub/api/dam/assets/22866367/master

OHM Paul, *Broken Promises of Privacy: Responding to the Surprising Failure Of Anonymization,* UCLA Law Review, Vol. 57, p. 1701, 2010.

PAPE Sebastian/ SERNA-OLVERA Jetzabel/ B. TESFAY Welderufael, *Why Open Data May Threaten Your Privacy,* septembre 2015.

PRASSER Fabian/ Kohlmayer Florian/A.KHUN Klaus, *The Importance of Context: Risk-based De-identification of Biomedical Data,* Methods of Information in Medicine 4/2016, p. 346ss.

Préposé fédéral à la protection des données, *Guide relative aux mesures techniques et organisationnelles de la protection des données*, du 15 janvier 2024.
(cité : PFPDT MTO 2024)

ROSENTHAL David/ STUDER Samira/ LOMBARD Alexandre, *La nouvelle loi sur la protection des données,* Jusletter du 16 novembre 2020, disponible sous :
https://www.rosenthal.ch/downloads/Rosenthal-Studer-Lombard-nouvelleLPD.pdf

Single Resolution Board, *Q&A sur la résolution,* 2018, disponible sous :
https://www.srb.europa.eu/system/files/media/document/2018%20resolution%20Q%26A%20%28FR%29.pdf

Stiftung Datenschutz, *Basic Rules for the Anonymisation of Personal Data (process management, evaluation and monitoring),* novembre 2022, disponible sous:
https://stiftungdatenschutz.org/fileadmin/Redaktion/Dokumente/Anonymisierung_personenbezogener_Daten/SDS_Basic_Rules_for_the_Anonymisation-Web-EN.pdf
(cite: Stiftung Datenschutz *Basic rules)*

SWEENEY Latanya, *Simple Demographics Often Identify People Uniquely*, Carnegie Mellon University, Data Privacy Working Paper 3. Pittsburgh 2000.

SWEENEY Latanya/ SAMARATI Pierangela, *Generalizing Data to Provide Anonymity when Disclosing Information,* mai 1998.

*Swissubase User Guide*, Version 2.1.0 – Mai 2022, disponible:
https://resources.swissubase.ch/wp-content/uploads/2022/05/SUB_User-Guide-update-May-2022.pdf
(cité : Swissubase *User Guide)*

ZEVENBERGEN Bendert/ BROWN Ian/ WRIGHT Joss/ ERDOS David, *Ethical Privacy Guidelines for Mobile Connectivity Measurements,* Oxford Internet Institute, University of Oxford, novembre 2013, disponible sous :
https://www.freehaven.net/anonbib/cache/ZevenbergenBrownWrightErdos2013.pdf

## Textes officiels

Masterplan Open Government Data 2024-2027 de l'OFS, 2023, disponible sous :
(cite: OFS Masterplan 2024-2027).

Message concernant la loi fédérale sur la protection des données, du 23 mars 1988, FF 1988 421.

Message concernant la loi fédérale sur la révision totale de la loi fédérale sur la protection des données et sur la modification d'autres lois fédérales, du 15 septembre 2017, FF 2017 6565.

Stratégie du Conseil fédéral en matière de libre accès aux données publiques en Suisse pour les années *2019 à 2023*, du 30 novembre 2018*,* FF 2019 855.

Stratégie du Conseil fédéral en matière de libre accès aux données publiques en Suisse pour les années 2014 à 2018, du 16 avril 2014, FF 2014 3347.

## Législation

Code civil suisse du 10 décembre 1907 (CC), RS 210.

Convention du 28 janvier 1981 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, RS 0.235.1

Loi fédérale du 19 décembre 1986 contre la concurrence déloyale (LCD), RS 241.

Loi fédérale du 30 septembre 2011 relative à la recherche sur l'être humain (LRH), RS 810.30.

Règlement (UE) du 15 juillet 2014 établissant des règles et une procédure uniforme pour la résolution des établissements de crédit et de certaines entreprises d'investissement dans le cadre d'un mécanisme de résolution unique et d'un Fonds de résolution bancaire unique, n. 806/2014.

Loi fédérale du 18 mars 2016 sur l'enregistrement des maladies oncologiques (LEMO), RS 818.33.

Règlement (UE) du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données personne et à la libre circulation de ces données (RGPD), n. 2016/679.

Ordonnance du 11 avril 2018 sur l'enregistrement des maladies oncologiques (OEMO), RS 818.331.

Loi fédérale du 25 septembre 2020 sur la protection des données (LPD), RS 235.1.

Ordonnance du 31 août 2022 sur la protection des données (OPDo), RS 235.11.

Loi fédérale du 17 mars 2023 sur l'utilisation de moyens électroniques pour l'exécution des tâches des autorités (LMETA), RS 172.019.

## Table des arrêts

ATF 136 II 508, jdT 2011 II p. 446

ATF 138 II 346, jdT 2013 I p. 71

HGer ZH, HG150170, 30.5.2017
Tribunal de l'Union européenne, Arrêt du 26 avril 2023, CRU c/ CEPD, T-557/20, EU:T:2023:219

CJUE, Arrêt du 20 décembre 2017, Peter Nowak c/ Data Protection Commissioner., C-434/16, EU:C:2017:994.

# Table des abréviations

| | |
|---|---|
| AEPD | *Agencia Española de Protección de Datos* (en françias : agence espagnole de protection des données) |
| AG | *Aktiengesellschaft* (en français : Société anonyme) |
| AJP | *Akutelle Juristiche Praxis* |
| aLPD | ancienne Loi fédérale sur la protection des données personnelles |
| art. | article |
| CEPD | Contrôleur européen de la protection des données |
| CHF | Francs suisses |
| CHIPS | Clearing House Interbank Paymens System |
| CJUE | Cour de justice de l'Union européenne |
| CR | Commentaire romand |
| CRU | Conseil de résolution unique |
| éd. | édition |
| édit./édits | éditeur/éditeurs |
| ENISA | *European Union Agency for Cybersecurity* (en français: Agence de l'Union européenne pour la cybersécurité) |
| FinCEN | Financial Crimes Enforcement Network |
| G29 | Groupe de travail Article 29 sur la protection des données |
| HIPAA | Health Insurance Portability and Accountability Act |
| ICO | *Information Commissioner's Office of United Kingdom* (en français: Commissaire à l'information du Royaume Uni) |
| IMDb | Internet Movie Database |
| IP | Internet Protocol (en français : protocole Internet) |
| IPC | Information and Privacy Commissioner of Ontario (en français: Commissaire à l'information et à la protection de la vie privée de l'Ontario) |
| ISO | International Organization for Standarization (en français : Organisation internationale de normalisation) |
| LPD | Loi fédérale sur la protection des données personnelles |
| LRH | Loi fédéral relative à la recherche sur l'être humain |
| MIT | Massachussetts Institute of Technology |
| N | Numéro |
| OFS | Office fédéral de la statistique |
| OFSP | Office fédéral de la santé publique |
| PFPDT | Préposé fédéral à la protection des données et à la transparence |
| PJA | Pratique Juridique Actuelle |
| RGPD | Règlement général sur la protection des données de l'Union européenne |

| | |
|---|---|
| RS | Recueil systématique suisse |
| ss | et suivant(e)s |
| SWIFT | Society for Worldwide Interbank Financial Telecommunication (en français: société de télécommunication financière interbancaire mondiale) |
| SZW-RSDA | *Schweizerische Zeitschrift für Wirtschafts- und Finanzmarktrecht* (en français : Revue suisse de droit des affaires et du marché financier) |
| TAF | Tribunal administratif fédéral |
| TF | Tribunal fédéral |
| TUE | Tribunal de l'Union européenne |
| UE | Union européenne |
| vs | versus<br>(en français : contre) |

# I. Introduction

En 2014, Chris W., un urbaniste et spécialiste en données, a obtenu des données sur les trajets de taxis de New York grâce à une demande fondée sur la *Freedom of Information Law*. Cette loi qui est fondée sur le droit à l'information, oblige les agences gouvernementales à transmettre leurs documents. Ces données comprenaient un grand nombre d'information, notamment l'horaire de prise en charge, les points de départ et d'arrivé, le type de paiement, les pourboires reçus ainsi que les numéros de licence des taxis. Afin de garantir le respect de la vie privée, la base de données avaient été « anonymisée », ou du moins c'est ce que l'agence l'ayant divulguée croyait[1].

Chris W. a décidé de rendre cette base de données publique sur Internet et c'est alors qu'un internaute de *Reddit*[2] s'est aperçu que les colonnes soi-disant anonymes avaient été chiffrées à travers le protocole MD5 qui est une fonction de hachage cryptographique. Une des propriétés des fonctions de hachage cryptographique est la résistance à la préimage, c'est-à-dire qu'il est extrêmement compliqué de remonter au message d'origine en connaissant uniquement l'empreinte numérique[3]. Cependant, si l'on connait quelques informations sur les caractéristiques de l'*input* initial ainsi que le type de fonction utilisée, cela devient alors plus facile. Il est par exemple possible de tester systématiquement toutes les combinaisons possibles jusqu'à trouver une correspondance avec l'empreinte. Or dans le cas d'espèce des informations sur l'*input* initial étaient bien connues. En effet, les informations qui avaient été « anonymisées » étaient les numéros de licence de taxi qui suivent un certain format structuré. Il a fallu moins d'une heure pour désidentifier l'intégralité du jeu de données. Une fois obtenu le numéro de licence, il a été facile de remonter à l'identité du propriétaire en effectuant une simple recherche *Google*. Mais les informations que l'on a pu apprendre ne s'arrête pas à l'identité du titulaire de la licence. La base de données contenait tous les paiement reçus, il a alors été possible de calculer le revenu brut mensuel de chaque taxiste. De plus, en recoupant les données sur les pourboires avec des photographies prises par des paparazzis, il a même été possible de déterminer quelle célébrité avait les poches cousues[4].

Cet exemple n'est qu'un parmi tant d'autres où un ensemble de données prétendument anonymisé a été publié, entraînant ainsi une atteinte à la vie privée[5]. La croissance exponentielle des flux informationnels[6] ainsi que l'accessibilité et le partage encouragé par des initiatives telles que *l'open data*[7], les lois sur la transparence, l'utilisation des réseaux sociaux, les technologies de *cloud computing*, mais également les avancées techniques de désidentification ou l'amélioration de la puissance de calcul rendent aujourd'hui extrêmement difficile l'anonymisation véritable d'un jeu de données. Les lois sur la protection des données, qui adoptent une approche dichotomique – soit les données sont anonymes et donc non soumises aux réglementations sur la protection des données, soit elles sont personnelles et donc régulées – sont confrontées à des défis majeurs[8]. Ce travail vise à analyser la tension entre le droit de la protection des données et les besoins toujours croissant de la société de connaître et analyser

---

[1] https://blogs.lse.ac.uk/impactofsocialsciences/2014/07/16/nyc-improperly-anonymized-taxi-logs-pandurangan/ (consulté le 22 mars 2024)

[2] *Reddit* est un site web communautaire américain de discussion et d'actualités sociales (source : Wikipedia).

[3] https://fr.wikipedia.org/wiki/Fonction_de_hachage_cryptographique (consulté le 7 juin 2024).

[4] *https://www.salingerprivacy.com.au/2015/04/19/bradley-coopers-taxi-ride-a-lesson-in-privacy-risk/* (consulté le 22 mars 2024)

[5] KAMATH, p. 1.

[6] FF 88.032 424

[7] Voir par exemple : FF 2014 3347.

[8] JOTTERAND, p. 3.

l’information. Pour ce faire, le travail a été divisé en trois parties. Dans la première, nous nous concentrerons sur la notion de donnée personnelle d’un point de vue légal, un concept essentiel mais quelque peu flou. Dans la deuxième partie, nous allons présenter différentes techniques d’anonymisation ainsi que quelques cas pratiques où l’anonymisation a échoué. Enfin, dans la troisième partie, nous analyserons divers scénarios où une divulgation de données peut se produire.

# II. La protection des données personnelles en Suisse

## A. Cadre légal

La protection des données personnelles est devenue un enjeu crucial dans notre société contemporaine, en raison notamment d'une multiplication de traitements de données dû à l'avènement des technologies de l'information[9].

En Suisse, le traitement de données personnelles effectué par des personnes privées ou par des organes fédéraux est régulé par la Loi fédérale sur la protection des données (LPD), adoptée pour la première fois en 1992 et ayant subi une révision complète récemment pour s'adapter aux évolutions technologiques[10]. Fédéralisme l'impose, il existe également 24 lois cantonales qui réglementent la protection des données (le Jura et Neuchâtel ayant adopté une Convention intercantonale commune)[11]. Cependant, ces lois ont un champ d'application différent, dans la mesure où elles réglementent exclusivement le traitement de données personnelles effectué par des autorités cantonales.

Contrairement à ce que son nom pourrait laisser entendre, la LPD ne vise pas à protéger les données personnelles en tant que telles. Son objectif est plutôt celui de protéger la personnalité et les droits fondamentaux des personnes concernées par le traitement de leurs données[12]. Pour ce faire, la loi établit aux art. 6 à 8 LPD des principes de base que tout traitement de données doit respecter pour être considéré comme licite.

Outre la loi fédérale et les lois cantonales, la Suisse est liée par la Convention 108 du Conseil de l'Europe[13], un instrument international conçu pour protéger les individus contre le traitement automatisé de leurs données personnelles. En ratifiant cette convention, ainsi que sa version modernisée, la Convention 108+, la Suisse s'est engagée à respecter ses dispositions[14].

A noter que le Règlement général sur la protection des données de l'Union européenne (RGPD) peut également trouver application en Suisse de par son article 3[15]. Compte tenu de l'influence significative du RGPD en Suisse ainsi que des similitudes avec la LPD, nous allons également nous y référer lorsque cela semblera pertinent.

En sus des législations citées ci-dessus, il existe également des réglementations sectorielles[16] fixant des exigences particulières en matière de traitement de données personnelles dans des domaines spécifiques comme la Loi fédérale relative à la recherche sur l'être humain (LRH) qui fixe des exigences en matière de traitements de données liés à des projets de recherche sur les maladies humaines et sur la structure et le fonctionnement du corps humain (Art. 2 al. 1 LRH)[17].

---

[9] FF 88.032 424

[10] FF 17.059 6795 ; Loi fédérale du 25 septembre 2020 sur la protection des données (LPD), RS 235.1.

[11] ERARD p.4.

[12] Art. 1 LPD.

[13] Convention du 28 janvier 1981 pour la protection des personnes à l'égard du traitement automatisé des données à caractère personnel, RS 0.235.1.

[14] https://www.edoeb.admin.ch/edoeb/fr/home/kurzmeldungen/2023/convention108.html (consulté le 23 mai 2024)

[15] JOTTERAND, p. 2.

[16] *Ibid.*

[17] Loi fédérale du 30 septembre 2011 relative à la recherche sur l'être humain (LRH), RS 810.30.

## B. Notion de donnée personnelle

La notion de donnée personnelle est au cœur du droit de la protection des données, car elle détermine le champ d'application de la LPD[18].

Les données personnelles sont définies à l'art. 5 let. a LPD comme étant *« toute information concernant une personne physique identifiée ou identifiable ».* Avant l'entrée en vigueur de la nouvelle LPD le 1er septembre 2023, la définition était identique, à l'exception de l'absence du mot « physique ». En effet, l'ancienne loi protégeait également les données des personnes morales (art. 3 let. a aLPD).

On retrouve des définitions très similaires sur le plan international. Tout d'abord, la Convention 108+ révisée définit les données à caractère personnel comme *« toute information relative à une personne physique identifiée ou identifiable »* (art. 2 let a Convention 108+). Tandis que le RGPD les décrit comme étant *« toute information se rapportant à une personne physique identifiée ou identifiable »* (art. 4 ch. 1 RGPD).

Il est reconnu tant au niveau suisse qu'européen une acceptation large de la donnée personnelle[19]. A titre d'exemple, dans l'Affaire C-434/16 (Nowak), la Cour de justice de l'Union européenne a considéré qu'une copie d'examen constituait une donnée personnelle et partant pouvait être consultée grâce au droit d'accès[20].

Les données personnelles peuvent en outre constituer des données sensibles, c'est-à-dire des données que le législateur considère être particulièrement dignes de protection parce que leur traitement présente un risque particulier d'atteinte à la personnalité et aux droits fondamentaux (notamment un risque de discrimination)[21]. Il s'agit de données *« sur les opinions ou les activités religieuses, philosophiques, politiques ou syndicales, les données sur la santé, la sphère intime ou l'origine raciale ou ethnique, les données génétiques, les données biométriques identifiant une personne physique de manière univoque, les données sur des poursuites ou sanctions pénales et administratives, les données sur des mesures d'aide sociale»* (art. 5 let. c LPD). Ces données sont soumises à des exigences supplémentaires en vertu de la loi[22]. Par exemple, pour pouvoir communiquer des données sensibles à des tiers, il est nécessaire de disposer d'un motif justificatif au sens de l'art. 31 LPD (consentement de la personne concernée, un intérêt privé ou public prépondérant ou une base légale).

L'on observe une approche dichotomique dans les lois de protection des données, où les données sont soit des données personnelles soit des données anonymes, et les lois ne s'appliquent généralement pas dans ce dernier cas. Malgré l'importance cruciale de la notion de donnée personnelle dans la mesure où elle détermine le champ d'application de la loi, ses contours semblent parfois manquer de clarté[23]. Nous allons maintenant décomposer sa définition afin d'identifier précisément où se situent les zones floues.

La notion de donnée personnelle peut se subdiviser en quatre composantes : une information, une personne physique, un lien et une possibilité d'identification[24].

### 1. Notion d'information

La notion d'information couvre un large spectre. Elle comprend tant des éléments de fait (par exemple une expérience professionnelle ou une maladie) que des jugements de valeurs (par

---

[18] CR LPD – MEIER/TSCHUMY, art. 5 N 18.

[19] *Ibid* N 19 ; WP 136 p. 4 ; MEIER N 422.

[20] CJUE, Arrêt du 20 décembre 2017, Peter Nowak c/ Data Protection Commissioner., C-434/16, EU:C:2017 :994, § 33ss.

[21] CR art. 5, p. livre 76 ph. 49

[22] Ibid, ph. 49.

[23] JOTTERAND, p. 3.

[24] JOTTERAND, p. 4.

exemple un pronostic de guérison)[25]. Elle peut revêtir plusieurs formes, comme des mots, des images, des signes, des sons ou une combinaison de ces éléments[26]. Le support de l'information peut aussi varier : document papier ou support informatique[27].
Le caractère confidentiel ou non d'une information ne joue aucun rôle au stade de sa définition. La même chose vaut pour le caractère exact ou moins de l'information. Cela découle notamment du droit de rectification prévu par la loi[28].

2. Une personne physique

Comme évoqué précédemment, la donnée personnelle doit concerner une personne physique. Les personnes morales ne sont plus protégées depuis l'entrée en vigueur de la nouvelle LPD le 1er septembre 2023. Ce faisant, la Suisse s'aligne avec l'Union européenne ainsi que la Convention 108+. A noter cependant que les personnes morales peuvent trouver protection contre un mauvaise utilisation de leurs données à travers d'autres dispositions de l'ordre juridique suisse, notamment l'art. 28 du Code civil ou la Loi sur la concurrence déloyale[29].

3. Un lien

L'information doit être *en lien* avec une personne physique. Il existe trois types de lien qui sont: un lien de **contenu**, un lien de **finalité** et un lien de **résultat**. Dans de nombreuses situations, ce lien est facile à établir parce qu'il est direct, c'est-à-dire que le contenu même de l'information se rapporte à une personne. On parle alors de lien de contenu. Par exemple, un dossier médical ou un *curriculum vitae,* de par leur conception même, sont liés à une personne[30].
Il y a d'autres situations où ce lien est moins évident, parce qu'il est indirect. Par exemple, des informations relatives à des objets peuvent constituer des données personnelles dans la mesure où elles fournissent des informations sur leur propriétaire ou toute autre personne exerçant une influence sur ces objets. De la même manière, des informations sur des événements peuvent nous en dire plus sur les personnes y ayant participé ou des informations sur un lieu peuvent nous en dire plus sur la personne qui s'y trouve[31]. La question qui se pose est alors de savoir dans quel cas ces informations *sont en lien* avec une personne ?
Premièrement, lorsqu'il existe un élément de **finalité**. C'est le cas lorsque les données sont utilisées ou pourraient être utilisées afin d'évaluer, de traiter ou d'influer sur une personne[32].
Deuxièmement, faute d'un élément de finalité ou de contenu, il peut y avoir un élément de **résultat.** On retiendra cet élément si des données sont susceptibles d'avoir un impact sur certains droits et intérêts d'une personne. A titre d'exemple, un système de localisation de taxis par satellite, qui a pour objectif de rendre le service plus efficace, ne présente ni un lien de contenu ni de finalité, dans la mesure où la localisation d'un taxi n'est pas directement liée à une personne et que le système n'a pas été mis en place pour évaluer ou influer sur une personne. Mais puisque ce système a pour résultat le contrôle des déplacements des chauffeurs, il s'agit d'une donnée personnelle[33].
A noter qu'une même donnée peut être une donnée personnelle pour plusieurs personnes. Pour reprendre l'exemple de l'Affaire C-434/16 (Nowak) mentionnée précédemment, la Cour de justice européenne a jugé que les annotations de l'examinateur constituaient des données

[25] MEIER, N 422.
[26] CR LPD – MEIER/TSCHUMY, art. 5 N 20.
[27] *Ibid.*
[28] MEIER, N 422.
[29] FF 2017 6565, p. 6632.
[30] WP 136 p. 10ss ; MEIER N 425.
[31] WP 136 p.10ss.
[32] *Ibid.*
[33] *Ibid.*

personnelles à la fois pour l'examinateur (dans la mesure où le contenu reflétait son avis) que pour le candidat évalué (puisque le contenu constitue un reflet de ses performances)[34].

4. Une personne identifiée ou identifiable

La personne est identifiée lorsqu'il ressort directement des informations détenues qu'il s'agit d'une personne déterminée et d'elle seule. C'est le cas d'une pièce d'identité par exemple[35]. Un prénom peut également suffire à identifier directement une personne, à condition que cette personne se distingue au sein d'un groupe dans un contexte donné[36]. On comprend alors que l'identification est étroitement liée au contexte. Par exemple, un nom commun ne sera pas suffisant pour identifier une personne au sein d'une population d'un pays, mais pourrait l'être pour identifier une personne dans une classe[37].
L'identification peut se faire sur la base d'un seul élément (par exemple « le Président des Etats-Unis ») ou sur la base d'une combinaison d'éléments[38].
Le RGPD à ce titre a été plus exhaustif que la LPD en fournissant dans sa définition même de donnée personnelle une liste exemplative d'identifiants potentiels : « un *nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale »* (art. 5 al. 1 RGPD).
La personne est considérée comme identifiable lorsque les informations disponibles, bien qu'elles ne permettent pas à elles seules une identification directe, peuvent permettre de reconnaître la personne en les combinant avec d'*autres « informations tirées des circonstances ou du contexte »*[39]. Le caractère identifiable est celui qui a suscité le plus de débat[40]. Tant la doctrine suisse qu'européenne a interprété de façon différente cette composante de la définition. On distingue généralement deux approches opposées : l'approche absolue et l'approche relative. Nous allons maintenant présenter ces deux approches ainsi que quelques cas de jurisprudence pour illustrer comment les tribunaux ont appréhendé la question.

5. Approche absolue *vs* relative du caractère identifiable

La divergence d'interprétation entre le courant absolu et le courant relatif a trait aux limites de ce qui doit être considéré comme identifiable. L'approche absolue considère que les données doivent être qualifiées de personnelles dès qu'une possibilité théorique d'identification existe. Cette possibilité doit exister tant du point de vue des <u>moyens</u> que des <u>acteurs</u>. Plus particulièrement :

- **Les moyens** : tous les moyens possibles pour identifier une personne doivent être pris en compte, indépendamment du coût ou de la faisabilité pratique ;
- **Les acteurs** : il suffit qu'une personne au moins puisse identifier l'individu et ce même si le détenteur actuel des données ne peut pas le faire[41].

L'approche relative, en revanche, soutient que les données ne doivent être considérées comme personnelles que si une identification est raisonnablement envisageable par le détenteur des données. Sous l'angle des deux composantes cela donne :

---

[34] CJUE, Arrêt du 20 décembre 2017, Peter Nowak c/ Data Protection Commissioner., C-434/16, EU:C:2017 :994, § 44.

[35] MEIER N 431.

[36] *Ibid.*

[37] WP 136 p. 14.

[38] MEIER N 433.

[39] CR LPD – MEIER/TSCHUMY, art. 5 N 23 ; Message LPD FF 2017 6639.

[40] JOTTERAND, p. 6.

[41] *Ibid.*

- **Les moyens** : seuls les moyens réalistiquement disponibles pour identifier une personne doivent être pris en compte ;
- **Les acteurs** : seul le point de vue du détenteur des données est pris en considération. Les données seront alors personnelles pour celui qui peut identifier l'individu, mais pas nécessairement pour celui qui les reçoit[42].

Le Conseil fédéral, dans son Message sur la nouvelle LPD s'est positionné pour une approche relative en indiquant que si les efforts nécessaires à identifier une personne sont tels que d'après le cours ordinaire des choses personne ne s'y attèlera, alors il ne faut pas parler de possibilité d'identification. Chaque situation doit être évaluée en tenant compte des moyens raisonnablement susceptibles d'être utilisés pour identifier la personne. Pour juger du caractère raisonnable de ces moyens, il faut prendre en compte l'ensemble des circonstances comme le coût, le temps nécessaire à l'identification, les technologies disponibles au moment du traitement ainsi que leur évolution et l'intérêt que peut représenter l'identification[43].
La conséquence de ce raisonnement est que les données anonymisées ne seront plus soumise à la LPD lorsqu'il est nécessaire de déployer des efforts considérables pour les réidentifier et qu'il est raisonnable de penser que personne ne s'y attèlera. La même règle vaut pour les données pseudonymisées[44].
Le caractère identifiable est donc un concept extrêmement dynamique. Plus la durée de conservation d'une information est longue, plus le risque d'identification devient élevé[45].
La doctrine juridique suisse semble également pencher majoritairement pour une approche relative[46]. Cependant, comme indiqué justement par JOTTERAND dans son article, il n'est pas toujours clair si les auteurs qui adoptent une approche relative le font pour les deux composantes (acteur et moyens) ou pour une seule[47]. En effet, une approche plus nuancée est envisageable, et c'est d'ailleurs ce vers quoi semblent s'orienter les tribunaux suisses. Nous allons maintenant examiner quelques cas de jurisprudence suisse et européenne.

*6.* L'arrêt Logistep *AG*

La société Logistep AG avait développé un logiciel qui recherchait des œuvres protégées par les droits d'auteur offertes sur des réseaux *peer-to-peer*[48]. En cas de téléchargement desdites œuvres, différentes informations, dont les adresses IP, étaient enregistrées dans une base de données afin de pouvoir dans un deuxième temps transmettre aux titulaires des droits d'auteurs ces informations. Les titulaires des droits d'auteurs déposaient ensuite une plainte pénale contre inconnu et apprenaient l'identité cachée derrière l'adresse IP grâce à leur droit de consulter le dossier. Une action civile en dommages et intérêts était ensuite entreprise[49].
Le Préposé fédéral à la protection des données (PFPDT) ayant pris connaissance de ce *business model,* a alors recommandé à la société d'arrêter immédiatement l'activité car elle contrevenait à la législation sur protection des données personnelles. Par courrier du 14 février 2008, Logistep AG a rejeté la recommandation du PFPDT. Ce dernier a alors porté l'affaire devant le Tribunal administratif fédéral (TAF) en concluant principalement à ce que soit ordonné à

---

[42] JOTTERAND, p. 6.

[43] Message LPD FF 2017 p. 6639.

[44] *Ibid*.

[45] WP 136 p. 16.

[46] MEIER N 445 ; JOTTERAND, p. 6 ; CR LPD – MEIER/TSCHUMY, art. 5 N 25.

[47] JOTTERAND, p. 6.

[48] Les réseaux pair-à-pair (de l'anglais : *peer-to-peer)* reposent sur un modèle d'échange en réseau ou chaque entité est à la fois client et serveur, par opposition à un modèle client-serveur (source : https://fr.wikipedia.org/wiki/Pair-%C3%A0-pair consulté le 7 juin 2024).

[49] ATF 136 II 508, JdT 2011 II p. 446.

Logistep AG de cesser immédiatement tout traitement de données personnelles, tant qu'il n'existerait pas de base légale pour ce type de traitement. Le TAF a rejeté la requête par jugement du 27 mai 2009 et annulé la recommandation du PFPDT[50].
Le PFPDT a alors saisi le Tribunal fédéral (TF) d'un recours en matière de droit public en date du 26 juin 2009. Logistep AG en sa défense soutenait que les adresses IP traitées par elle ne constituaient pas des données personnelles, puisqu'elle n'était pas en mesure d'identifier les personnes qui se cachent derrière ces adresses[51].
Le TF a débuté son argumentation en définissant ce qu'est une adresse IP : *« il s'agit d'un paramètre numérique qui permet d'identifier un domaine Internet composé notamment d'ordinateurs des usagers qui participent aux relations de communication sur ce réseau ».* Un ordinateur qui se connecte à internet sera alors identifié par l'adresse IP. Cette adresse peut être statique ou bien dynamique. On parle d'adresse *statique* lorsqu'elle est attribuée de manière fixe à un ordinateur. En revanche, une adresse est dite *dynamique* lorsqu'elle est attribuée de manière temporaire par le fournisseur d'accès Internet et qu'à chaque connexion l'utilisateur reçoit une nouvelle adresse. L'identification du titulaire d'une adresse IP dynamique est plus complexe que celle du titulaire d'une adresse statique. En effet pour pouvoir identifier une adresse IP dynamique il est nécessaire d'obtenir l'assistance du fournisseur d'accès Internet[52].
Ensuite, le TF poursuit en indiquant que la possibilité d'identification s'examine du point de vue du détenteur de l'information. Cependant, elle continue son raisonnement en indiquant qu'en cas de transmission d'informations, il suffit que le destinataire soit en mesure d'identifier la personne concernée pour qu'on soit en présence de données personnelles. Donc, dans le cas d'espèce, même si Logistep AG n'est pas en mesure d'identifier les individus qui se cachent derrière les adresses IP traite de données personnelles dans la mesure où elle transmet des informations à des individus qui, en déposant une plainte pénale, seront en mesure d'identifier les personnes concernées[53].
S'agissant des moyens pour réidentifier les personnes, le TF ajoute que l'on ne peut pas affirmer qu'ils sont tellement importants que, d'après le cours ordinaire des choses, personne ne les mettra en œuvre. Au contraire, le modèle d'affaire de l'intimé repose exactement sur cette possibilité[54].
Il ressort de cet arrêt que :

1) Les données sont personnelles si le détenteur peut identifier les individus, et ce même lorsqu'il doit recourir à l'assistance d'un tiers ;
2) Les données sont également considérées comme personnelles lorsque le détenteur ne peut pas réidentifier les individus, mais partage les données avec un tierce personne qui le peut[55].

Il est important de mentionner que cet arrêt a été interprété de manière contradictoire par la doctrine. Certains auteurs ont affirmé que le TF confirmait de par cet arrêt une approche relative[56]. D'autres ont perçu dans le raisonnement du TF une approche absolue[57], tandis que d'autres encore considèrent que le TF n'a suivi ni l'une ni l'autre des approche, préférant ainsi une solution plus nuancée[58].

---

[50] ATF 136 II 508, JdT 2011 II p. 446.

[51] *Ibid.*

[52] *Id.,* c. 3.3.

[53] *Id.,* c. 3.4.

[54] *Id.,* c. 3.5.

[55] JOTTERAND, p. 15.

[56] MEIER, Commentaire d'arrêt JdT 2011 II p. 446, point c, ERARD (code), p. 5.

[57] ROSENTHAL/ STUDER / LOMBARD, p. 9.

[58] JOTTERAND, p. 14.

### 7. L'arrêt *Google Street View*

Google a lancé le service « *Street View* » pour la Suisse en août 2009. L'enregistrement des rues a été exécuté dès mars 2009 au moyen de véhicules spécialement conçus à cet effet. Les visages des personnes ainsi que les plaques d'immatriculation des véhicules pris en image ont été automatiquement floutés. Des plaintes ont cependant été déposées auprès du PFPDT par des personnes qui se sentaient atteintes dans leur personnalité[59].

Le PFPDT a estimé que le floutage automatique des images par un logiciel était insuffisant car seul une partie des visages et des plaques d'immatriculation avaient été floutées. Le PFPDT a alors émis des recommandations à Google. Ce dernier les ayant rejetées, le PFPDT a alors ouvert action devant le TAF en soutenant notamment les conclusions suivantes :

- Google doit garantir que toutes les images soient totalement anonymisées avant de les publier ;
- Pour certains lieux sensibles, comme les foyers d'accueil pour femmes ou les hôpitaux, une précaution particulière doit être prise pour garantir l'anonymat des personnes ;
- Les espaces privés (cours fermées, jardins, etc.) ne doivent pas faire l'objet de prises de vue et les images déjà prises doivent être retirées ;
- Google doit informer la population concernée à l'avance lorsqu'il souhaite prendre de nouvelles images[60].

Le TAF a en grande partie confirmé les demandes du PFPDT. Google[61] a alors interjeté recours en matière publique auprès du TF le 19 mai 2011[62].

Dans son recours, Google a soutenu que les images ne constituaient pas des données personnelles[63].

Le TF a alors indiqué que les données brutes (celles qui n'ont pas encore été anonymisées) constituaient sans aucun doute des données personnelles. En outre, il a reconnu que même en cas de floutage automatique, une identification restait parfois possible. Premièrement, c'était le cas lorsque le floutage se révélait défectueux (par exemple un floutage partiel ou pas de floutage du tout). Deuxièmement, c'était le cas lorsque, alors même que l'image avait été correctement floutée, le lieu et le contexte de la prise de vue permettaient quand-même de reconnaître l'individu. Par exemple, pour des personnes photographiées dans leur environnement habituel, une vraisemblance d'identification par des connaissances ou des voisins resterait toujours possible[64].

Le TF a ensuite ajouté que des données personnelles pouvaient être présentes sur des images même lorsqu'aucune personne n'y été représentée. Par exemple, un véhicule pourrait permettre un rapprochement entre son titulaire et l'adresse de son domicile[65].

Des images de jardins, cours, balcons privés et façades de maisons avec une vue sur des locaux d'habitation pourraient également, selon le contexte, constituer des données personnelles[66].

S'agissant du caractère identifiable, le TF a reconfirmé son raisonnement de l'arrêt Logistep AG : *« la question doit être résolue en fonction du cas concret, où il convient de tenir particulièrement compte des possibilités offertes par la technique, à l'exemple des outils de recherche disponibles sur Internet. Le coût objectivement nécessaire pour rattacher une*

---

[59] ATF 138 II 346, JdT 2013 I p. 71

[60] *Ibid.*

[61] Le recours a été interjeté par *Google Inc.* et sa filiale *Google Switzerland S.à.r.l.*

[62] ATF 138 II 346, JdT 2013 I p. 71

[63] *Id.* c. 6.

[64] *Id.* c. 6.2 et c. 6.3.

[65] *Ibid*.

[66] *Ibid*.

*information à une personne déterminée n'est pas le seul élément pertinent ; il faut également prendre en considération l'intérêt à l'identification que possède un tiers ou toute personne procédant au traitement des données. (..) Il convient d'examiner du point de vue du possesseur d'une information si celle-ci peut être mise en relation avec une personne sur la base d'éléments supplémentaires, de sorte qu'elle se rapportera à une personne identifiable. Dans le cas d'un transfert d'informations, il suffit que le destinataire puisse identifier la personne concernée. Il suffit en outre qu'une partie des informations enregistrées permette une identification »*[67].

Le TF a reconnu cependant la difficulté que représente l'anonymisation de l'intégralité des images. Il admet donc une marge d'erreur de 1% lors du floutage automatique. Cette marge d'erreur est soumise notamment aux conditions suivantes :

- Le floutage automatique doit être adapté régulièrement à l'état de la technique ;
- A proximité des établissement sensibles, comme les prisons ou les hôpitaux, une anonymisation complète des personnes, ainsi que des signes distinctifs, doit être effectuée avant publication[68].

Le TF a donc confirmé l'arrêt Logistep AG. Cependant, la situation diffère ici car le destinataire des données n'est pas une personne spécifique, mais potentiellement tous ceux qui ont accès à *Google Street View.* La définition de donnée personnelle prend alors toute son ampleur, car même des données floutées ou des images de jardins privés peuvent constituer des données personnelles.

## C. Transfert de données pseudonymisées

L'arrêt Logistep AG aborde la situation où c'est le détenteur des données qui n'est pas en mesure d'identifier les personnes tandis que le récipiendaire en est capable. Que se passe-t-il lorsque c'est le destinataire des données qui est incapable d'identifier les personnes concernées (parce qu'elles ont été pseudonymisées et qu'il ne dispose pas de la clé déchiffrement par exemple) ?

La doctrine juridique distingue généralement l'anonymisation de la pseudonymisation. L'*anonymisation* rend impossible (ou très difficile) le rattachement des données à une personne, même pour l'auteur du traitement. Tandis que la *pseudonymisation* consiste à remplacer des identifiants directs par un pseudonyme, tout en conservant une table de correspondance permettant de rattacher et donc réidentifier la personne. Doit-on considérer qu'il s'agit de données personnelles même pour celui qui n'aurait pas accès à cette table de correspondance ? La question est controversée en Suisse et cela découle du débat autour de l'approche relative et absolue[69]. Le *Handelsgericht* zurichois s'est penché sur la question et sa décision a fait l'objet d'un recours devant le TF, comme nous allons le voir dans la section suivante.

### 1. Transfert de données clients pseudonymisées vers les Etats-Unis

Dans le contexte du *Joint Statement Agreement* de 2013 visant à mettre fin au contentieux fiscal entre les Etats-Unis et les banques suisses, une banque a décidé de participer au programme du Département américain de justice et de l'administration fiscale américaine. Conformément au chiffre II.D.2 de ce programme, la banque était tenue de transmettre une liste d'informations (appelé aussi liste *Leaver*). Parmi les informations à fournir, il y avait : le nombre total de comptes détenus par des américains, la valeur maximale en dollars pour chaque compte, le nombre de personnes américaines affiliées à chaque compte, le nom du gestionnaire de relations, la nature de la relation entre le compte et la personne américaine. Le nom des clients

[67] ATF 138 II 346, JdT 2013 I p. 71, c. 6.1.

[68] *Id.,* c. 10.7.

[69] JOTTERAND, p. 17.

ainsi que les numéros de comptes étaient quant à eux remplacés par des pseudonymes. La table de correspondance permettant de relier les pseudonymes aux clients n'était pas transmise aux autorités américaines[70].
Deux clients de la banque ont introduit une action devant le Tribunal de commerce du canton de Zürich (*Handelsgericht*) en demandant que soit interdit à la défenderesse (la banque) de transmettre leurs données personnelles à l'étranger, y compris des données pseudonymisées ou cryptées. S'agissant des données personnelles déjà en possession des autorités américaines, ils demandaient à ce qu'elles soient détruites[71].
Le *Handelsgericht* a reconnu que des données pseudonymisées envoyées à un tiers (ci-après : récipiendaire) constituent des données anonymes pour ce dernier lorsqu'il ne dispose pas des moyens pour réidentifier les personnes. Cependant, il appartient à celui qui envoie les données (l'exportateur) de prouver qu'il a pris les mesures nécessaires pour empêcher la réidentification. Or, en l'espèce, la banque n'avait pas assez prouvé qu'elle avait pris les mesures nécessaires pour empêcher aux autorités américaines de réidentifier les clients[72]. Par jugement du 30 mai 2017, le Handelsgericht a alors admis la demande principale des requérants[73].
La banque a alors formé recours en matière civile auprès du TF en demandant l'annulation du jugement.
Le TF devait alors déterminer si les données mentionnées dans la liste II.D.2 constituaient des données personnelles au sens de la LPD, notamment au regard des mesures de pseudonymisation prises par la banque[74].
Le TF a commencé son raisonnement en reprenant la définition usuelle de données personnelles, à savoir « *toutes ces informations qui se rapportent à une personne identifiée ou identifiable. Une personne est identifiée lorsqu'il ressort de l'information elle-même qu'il s'agit précisément de cette personne. Une personne est identifiable lorsqu'il est possible de déduire son identité par recoupement d'informations. Pour être identifiable, toute possibilité théorique d'identification n'est toutefois pas suffisante. Si l'effort à fournir est tel qu'il ne faut pas s'attendre, selon l'expérience générale de la vie, à ce qu'une personne intéressée s'en charge, il n'y a pas d'identifiabilité. (...) Ce qui est important, ce n'est pas seulement l'effort objectivement nécessaire pour pouvoir attribuer une information déterminée à une personne, mais aussi l'intérêt qu'a le responsable du traitement des données ou un tiers à l'identification»*[75].
Les données initiales sur lesquelles se fondent la liste *Leaver* sont sans aucun doute des données personnelles. En pseudonymisant ces données (numéro de compte et nom du titulaire), le recourant effectue un traitement de données personnelles qui tombe sous le coup de la LPD. Le résultat de ce traitement -le pseudonyme - n'est plus protégé par la loi. Lorsque la banque veut transmettre ce résultat, elle se prévaut d'une exception (l'exception au principe de ne pas transmettre des données personnelles de clients à des tiers). Cette exception est justifiée par des mesures de pseudonymisation qui ont été prises par la banque. Le TF confirme alors que l'instance précédente a donc considéré à juste titre que c'est à la banque de prouver l'efficacité des mesures prises[76].
Parmi les mesures prises par la banque : les données de paiement ont été agrées sur un mois. Ce procédé parait propre à empêcher l'identification des titulaires de compte. Le nom et le

[70] JACOT-GUILLARMOD/ HIRSCH, p. 1, 4A_365/2017, A.

[71] 4A_365/2017, B.

[72] 4A_365/2017, c. 5.1.2.

[73] *Id.,* B.

[74] JACOT-GUILLARMOD/HIRSCH, p. 9.

[75] 4A_365/2017, c. 5.

[76] *Id.,* c. 5.2.2.

numéro de compte ont été remplacés par un pseudonyme, ce qui constitue également une mesure appropriée. Cependant, le nom du gestionnaire de relations ainsi que d'autres données de la liste des départs sont fournis en clair et constituent un point de rattachement important permettant de déterminer l'identité. Pour les ayants droit économiques, le lieu de naissance et de résidence est également transmis. Il faut donc partir du principe qu'une identification est déjà possible avec les données de la liste *Leaver*[77].
Au regard de ce qui précède, le Tribunal fédéral a confirmé l'arrêt du Handelsgericht et rejeté le recours[78].
D'après certains auteurs, le TF aurait ici confirmé l'approche relative en rejetant le recours de la banque. Appliquer une approche relative au transfert de données pseudonymisées aurait l'avantage de faciliter l'exportation des données, même si cela serait tempéré par la répartition du fardeau de la preuve[79].
D'autres auteurs considèrent en revanche que le TF s'est contenté de confirmer que les mesures de pseudonymisation n'avaient pas été suffisamment prouvées sans aborder la question de l'identifiabilité[80].
La question reste donc assez controversée en Suisse et sera, on l'espère, clarifiée à l'avenir par une nouvelle décision.

2. Qu'en est-il en Europe ?

Tout comme en Suisse, la question du caractère identifiable d'une information est très débattue au sein de l'Union européenne. Dans l'arrêt de la CJUE du 19 octobre 2016 (Breyer c. Allemagne) il était question de savoir si des adresses IP constituent des données personnelles. La CJUE a rejoint l'analyse du TF dans l'arrêt Logistep en indiquant que les adresses IP dynamiques pouvaient être considérées comme des données personnelles si le responsable du traitement, ici les autorités publiques allemandes, disposaient de moyens légaux pour associer les adresses à une personne physique[81]. Un autre arrêt tout récent vient apporter des clarifications supplémentaires sur la question du transfert de données pseudonimisées à un tiers. Nous allons le résumer dans la section suivante.

3. Arrêt du Tribunal de l'Union européenne du 26 avril 2023 (CRU/CEPD)

En 2017, une banque espagnole, la *Banco Popular Español*, rencontre des difficultés financières menaçant ainsi la stabilité du système financier. C'est pourquoi le Conseil de résolution unique (CRU), une autorité européenne visant à promouvoir la stabilité financière en Europe, intervient dans l'affaire en décidant de soumettre la banque espagnole à une procédure de résolution en application du règlement (UE) n° 806/2014[82]. La procédure de résolution remplace une liquidation classique et est mise en place lorsque l'intérêt public impose d'assurer la continuité des fonctions financières et économiques essentielles[83]. Généralement, dans une procédure de de résolution, ce sont les actionnaires et créanciers à subir le plus de pertes. C'est pourquoi Deloitte, une société d'audit et de conseil, a été mandaté afin de valoriser la différence

[77] 4A_365/2017*,* c. 5.1.1, c. 5.3.2.

[78] *Id.,* c. 8.

[79] JACOT-GUILLARMOD, p. 3.

[80] JOTTERAND, p. 18.

[81] *Id.,* p. 25.

[82] Règlement (UE) du 15 juillet 2014 établissant des règles et une procédure uniforme pour la résolution des établissements de crédit et de certaines entreprises d'investissement dns le cadre d'un mécanisme de résolution unique et d'un Fonds de résolution bancaire unique, n. 806/2014 ; TUE, arrêt du 26.04.2023, CRU contre CEPD, T-557/20.

[83] Single resolution board, *Q&A sur la résolution*, p. 2.

de traitement, c'est-à-dire déterminer si les actionnaires et créanciers auraient subi moins de pertes si la banque avait fait l'objet d'une procédure d'insolvabilité classique[84].
Les actionnaires et créanciers ont été invités à soumettre des commentaires au CRU afin de contribuer à l'évaluation. Dans le but de garantir leur confidentialité, les commentaires ont été pseudonymisés avant d'être envoyés à Deloitte. Plus particulièrement, ils ont été filtrés, catégorisés et agrégés (lorsque des commentaires se ressemblaient, on en tenait un seul). En plus, les commentaires portaient un code alphanumérique et le CRU était la seule à pouvoir relier ce code aux données originales. Le code alphanumérique était nécessaire pour permettre de vérifier *à posteriori* que chaque commentaire avait été traité et dûment pris en compte. Deloitte n'avait donc aucun moyen de relier les commentaires aux identités des clients de la banque[85].
Cinq réclamations ont été déposées auprès du Contrôleur européen de la protection des données (CEPD) par des clients de la banque se plaignant de n'avoir pas été informé de la transmission de leurs données personnelles à Deloitte ainsi qu'à la *Banco Santander* (la nouvelle banque qui a racheté l'ancienne en difficulté pour un euro symbolique). En effet d'après l'art. 15 paragraphe 1, sous d), du règlement 2018/1725, le CRU aurait eu l'obligation d'informer les personnes concernées sur les destinataires des données à caractères personnel. La CEPD a alors adopté une décision rappelant à l'ordre le CRU pour avoir violé le règlement[86].
Le CRU a recouru alors auprès du Tribunal de l'Union européenne (TUE) en demandant l'annulation de la décision de la CEPD. A l'appui de son recours, le CRU indiquait notamment que les informations transmises à Deloitte ne constituaient pas des données à caractère personnel. Il soutenait que l'objectif se cachant derrière l'analyse des commentaires était d'évaluer des arguments de fait et de droit et non d'évaluer la personnalité des clients. Leur identité n'étaient pas pertinentes aux fins d'évaluer leurs commentaires. En revanche, le CEPD faisait valoir que le contenu des commentaires des actionnaires et des créanciers était une information les « concernant », puisque leurs réponses contenaient et reflétaient leur point de vue personnel. D'ailleurs, le fait que les réclamants aient exprimé des points de vue semblables, mais non identiques, à ceux d'autres participants ne signifiaient pas que leurs réponses ne reflétaient pas leur propre opinion[87].
Premièrement, le TUE devait donc décider si les informations en question « se rapportaient » à une personne physique ou pas. Il a commencé son raisonnement en rappelant une ancienne jurisprudence de la CJUE (l'affaire Nowak, C-434/16, EU:C:2017:994) dans laquelle la condition du lien entre donnée et personne avait été retenue parce que l'un des 3 critères (contenu, finalité ou résultat) était présent. Or, la CEPD n'avait pas analysé ces aspects, ni le contenu des commentaires produits. Il a même affirmé que d'après lui, toute opinion personnelle constituait une donnée à caractère personnel. Or, d'après le TUE cet examen est essentiel pour pouvoir déterminer le lien[88].
S'agissant du caractère identifié ou identifiable des données : le CRU a fait valoir que faute d'avoir transmis les informations permettant la réidentification des auteurs des commentaires, la simple transmission des codes alphanumériques constituait une transmission de données anonymes. La CRU soulignait également que le règlement ainsi que la jurisprudence de la Cour exigeaient qu'une évaluation du risque de réidentification soit faite pour décider si des données pseudonymisées envoyé à un tiers, sont anonymes ou pas pour ce tiers.
En revanche, la CEPD a considéré que « des données ''pseudonymisées'' le resteraient même

[84] TUE, arrêt du 26.04.2023, CRU contre CEPD, T-557/20.
[85] *Ibid.*
[86] *Ibid.*
[87] *Ibid.*
[88] *Ibid.*

lorsqu'elles seraient transmises à un tiers qui ne dispose pas des informations supplémentaires »[89].
Deuxièmement, le TUE devait se pencher sur la question de savoir si les données transmises à Deloitte étaient des informations se rapportant à une personne « identifiable ». Le Tribunal a commencé par rappeler l'arrêt Breyer C-582/14, EU:C:2016:779 où la CJUE devait déterminer si une adresse IP dynamique constituait une donnée personnelle à l'égard du fournisseur de services en ligne qui l'avait enregistré. Dans cet arrêt, la Cour avait jugé que le fait que les informations supplémentaires nécessaires pour identifier l'utilisateur d'un site Internet étaient détenues non pas par le fournisseur de services de médias en ligne, mais par le fournisseur d'accès à Internet de cet utilisateur n'apparaissait ainsi pas de nature à exclure que les adresses IP dynamiques enregistrées par le fournisseur de services de médias en ligne constituaient, pour celui-ci, des données à caractère personnel. Cependant, il ressort également de l'arrêt, Breyer (C-582/14, EU:C:2016:779), que la possibilité de réidentification doit s'analyser du point de vue du destinataire. La CEPD n'a pas examiné si les auteurs des commentaires transmises à Deloitte étaient directement identifiables par Deloitte, ou si elle possédait les moyens légaux et pratiques pour accéder aux informations supplémentaires nécessaires à la réidentification. En l'absence de cette analyse, le TUE annule la décision du CEPD[90].
A noter que quand bien même cet arrêt concerne le règlement (UE) 2018/1725 et non directement le RGPD, les conclusions semblent transposables au RGPD de par la similarité des dispositions. Cette décision a le mérite de clarifier la situation entourant la notion de personne « identifiable ». Elle confirme l'arrêt « Breyer » et s'éloigne donc d'une approche absolue pour adopter une approche fondée sur le risque. Elle rapproche également la conception européenne de celle en vigueur en Suisse, où l'analyse de l'identifiabilité doit également se faire selon l'environnement du détenteur des données[91].

## D. Conclusion intermédiaire

La LPD fournit une définition extrêmement large de la donnée personnelle, qui s'aligne sur celle du RGPD. Cette définition inclut toutes les informations qui se rapportent à une personne physique identifiée ou identifiable. Les identifiants directs, tels que les noms ou l'adresse, sont clairement considérés des données personnelles.
Le défi majeur réside dans la notion du caractère identifiable. Une question se pose alors : à partir de quel moment un ensemble d'identifiants indirects constitue-t-ils des données personnelles ? Deux théories principales s'opposent alors dans la doctrine. L'approche absolue qui considère qu'une personne est identifiable dès le moment où un acteur dispose des moyens pour l'identifier. L'approche relative en revanche considère qu'il faut évaluer le caractère identifiable du point de vue de l'acteur qui possède les données.
A la lumière des cas de jurisprudence exposé ci-dessus, il semblerait que le TF ne s'aligne jamais strictement sur l'une ou l'autre approche, mais adopte une approche contextuelle en fonction des spécificités de chaque cas. Le monde juridique semble être de plus en plus conscient qu'une anonymisation irréversible est difficilement réalisable. C'est pourquoi l'approche contextuelle basée sur une sorte d'évaluation des risques semble à mon avis être la meilleure solution.
Nous allons maintenant entrer dans un chapitre plus technique en présentant les techniques d'anonymisation.

[89] TUE, arrêt du 26.04.2023, CRU contre CEPD, T-557/20.

[90] *Ibid.*

[91] JOTTERAND (arrêt TUE), p. 5.

# III. Différentes techniques d'anonymisation[92]

## A. Introduction

Pendant longtemps, les pratiques d'anonymisation des données se sont basées sur des méthodes relativement simples : suppression des identifiants directs comme les noms et les numéros de sécurité sociale. L'objectif était de protéger la vie privée des individus. Cependant, tant la pratique que les recherches ont révélé que ces techniques de désidentification sont souvent insuffisantes. Les possibilités de réidentification ont augmenté grâce à des algorithmes toujours plus sophistiqués, une puissance de calcul en constante augmentation ainsi qu'une disponibilité croissante d'informations auxiliaires (notamment grâce au Big Data)[93]. Il semblerait donc que, comme le dit Paul OHM dans son article sur les échecs de l'anonymisation : « *Data can be either useful or perfectly anonymous but never both* » [94]. Mais, *est-ce vraiment le cas ?* La tension entre utilité des données et protection des données est certainement aux cœurs du débat. Nous allons tenter d'y répondre en présentant de manière générale les différentes techniques d'anonymisation, tout en reconnaissant les limites de notre perspective de juriste.

Les autorités de protection des données européennes établissent généralement trois risques de réidentification sur un ensemble de données anonymisé :

1. **L'individualisation** : est-il possible d'isoler un individu dans l'ensemble de données ?
2. **La corrélation** : est-il possible de relier entre elles deux informations se rapportant à la même personne ou à un même groupe de personnes dans la même base de données ou bien dans deux bases de données distinctes ?
3. **L'inférence** : est-il possible de déduire des nouvelles informations sur un individu à partir de la base de données ? [95]

Une solution qui permettrait de répondre par la négative à ces trois questions offrirait des garanties fiables contre les tentatives de réidentification[96].

Dans la présente section, nous allons présenter des techniques d'anonymisation qui peuvent être utilisées sur des ensembles de données structurées, le plus souvent représentées sous forme de tableau. Pour anonymiser des ensembles de données plus complexes tels que des images ou du son, d'autres techniques existent mais ne seront pas traitées ici.

Outre à permettre dans certains cas de sortir du champ d'application de la LPD, les techniques d'anonymisation constituent une bonne solution pour garantir le principe de proportionnalité et de conservation ainsi que le principe de sécurité des données. En effet, dans la mesure où l'anonymisation est assimilée à une destruction, il s'agit d'une bonne solution afin de ne pas garder des données personnelles plus longtemps que nécessaire (art. 6 ch. 4 LPD). En outre, d'un point de vue sécuritaire, si les données devaient tomber dans des mauvaises mains, l'information dévoilée serait moins précise et moins sensible[97].

### 1. Vocabulaire

Le vocabulaire est repris de celui utilisé par le Groupe de travail Article 29 sur la protection des données. Le tableau suivant (tableau 1) constitue un ensemble de données sous forme de tableau. Il est composé de deux *enregistrements*, un relatif à chaque individu. Chaque enregistrement se compose d'une série de *valeurs* rattachées à des *attributs*. Les attributs sont le « nom », la « profession » et l'« âge ». Les valeurs sont les entrées des attributs. Pour le

[92] Le terme « anonymisation » a été repris du document WP 216, cependant le terme « désidentification » sera également utilisé en tant que synonyme.

[93] MEIER (Big Data), p. 47ss.

[94] OHM, p. 1703.

[95] WP 216 p. 13.

[96] WP 216 p. 13.

[97] PFPDT MTO 2024 p. 25.

premier enregistrement il s'agit donc de « Barbara », « Employé de commerce » et « 22 ». Il est également possible de concevoir un tableau où plusieurs enregistrements se rapportent à un même individu. Un *attaquant* est un tiers, autre que le responsable du traitement ou le sous-traitant, qui accède aux données de manière accidentelle ou intentionnelle.

| **Nom** | **Profession** | **Age** |
|---|---|---|
| Barbara | Employé de commerce | 22 |
| Luca | Informaticien | 37 |

**Tableau 1**

Les combinaisons d'attributs se rapportant à un même individu peuvent constituer des *quasi-identifiants.* Les *classes d'équivalence* sont des groupes de lignes présentant des quasi-identifiants identiques[98].

a) Quasi-identifiants

Introduit dans les travaux sur la réidentification de Latanya SWEENEY, ce concept s'est développé suite à une constatation assez troublante : les détenteurs de données libéraient souvent des bases de données en supprimant les identifiants directs tels que le nom, l'adresse ou le numéro de sécurité sociale. Ce faisant, ils pensaient avoir assuré la confidentialité des individus. Cependant, même en présence d'informations démographiques très générales voir des données matérielles, il est possible de réidentifier des individus en combinant ces informations avec des sources externes. Une étude menée aux Etats-Unis dans les années 2000 a montré qu'une combinaison de trois informations, à savoir le sexe, le code postal et la date de naissance, combiné avec une liste électorale accessible au public, permettait d'identifier 87% de la population des Etats-Unis[99].

On peut alors définir les quasi-identifiants comme un ou plusieurs attributs au sein d'un jeu de données qui, bien qu'ils ne soient pas directement identifiants, peuvent être utilisés conjointement avec des sources de données externes pour réidentifier des individus. Pour protéger la confidentialité des données, il est donc essentiel de reconnaître et gérer les quasi-identifiants. C'est à cette problématique que le k-anonymat tente de répondre, comme nous allons le voir plus tard.

2. Résultat d'une technique d'anonymisation

Le résultat qu'on obtient après avoir appliqué une technique d'anonymisation peut être de deux ordres :

- **Macro-données** (ou données agrégées) : ce sont des données qui donnent des informations sur un groupe d'individus dans son ensemble ;
- **Micro-données** : ce sont des données qui apparaissent sous la forme d'enregistrements individuels[100].

Dans ce sens, les données sous forme agrégée sont beaucoup plus robustes face à des tentatives de désidentification mais présentent une utilité limitée à certains usages[101].

## B. Types d'attaquants

Avant d'aborder les techniques d'anonymisation, il semble pertinent de présenter les types d'attaquants possibles. Généralement, l'on distingue trois modèles d'attaquants selon leurs motivations et leurs connaissances[102] :

- **Le modèle du procureur** : le but est de retrouver l'enregistrement spécifique relatif à un individu *i* tout en sachant que *i* se trouve dans la base de données. Soit

[98] WP 216 p. 13.

[99] SWEENEY, p. 17, PIERANGELA/SWEENEY, p. 3.

[100] ISO/IEC 20889:2018, p. 20.

[101] *Ibid.*

[102] PRASSER/ KOHLMAYER/ A. KUHN, p. 3.

*connaissance,* la probabilité que l'individu *i* est dans la base de données, et *corrélation,* la probabilité d'associer *i* à l'enregistrement correct. La probabilité de réussir à associer correctement cet individu à son enregistrement, en l'absence d'informations supplémentaires, est alors déterminée par l'inverse du nombre total d'individus présents dans l'ensemble de données *s* :

$$P\ (corrélation\ |connaissance) = (\frac{1}{s})\ ;$$

- **Le modèle du journaliste** : le but est de trouver qui se cache derrière un enregistrement spécifique. Si l'attaquant n'a aucune connaissance préalable de la présence de l'individu *i* dans l'ensemble de données (*connaissance*). La probabilité que l'individu *i* soit présent dans l'ensemble de données *s* est donné par la fraction de la taille de l'ensemble de données par rapport à la taille de la population totale (p) :

$$P\ (connaissance) = \frac{s}{p}$$

La probabilité de corrélation, sans information préalable est alors :

$$P(connaissance)\ x\ P\ (corrélation\ |connaissance) = \frac{s}{p}\ x\ (\frac{1}{s})\ ;$$

- **Le modèle du marketeur** : l'attaquant n'a aucune connaissance préalable et son but, contrairement au modèle précédent, n'est pas d'identifier un individu précis mais il souhaite plutôt identifier le plus grand nombre d'individus présents dans l'ensemble de données. La probabilité liée au modèle du marketeur est alors une moyenne des risques de réidentification de tous les enregistrements[103].

Identifier des adversaires plausibles est une étape essentielle dans l'évaluation du risque de réidentification. En effet, selon le type d'adversaire, les mesures d'anonymisation doivent être plus ou moins robustes[104].
Protéger un jeu de données contre une attaque basée sur le modèle du procureur va également protéger contre des attaques de type journaliste et de type marketeur[105].

## C. Outils statistiques

### 1. Échantillonnage

L'échantillonnage est une technique d'analyse statistique consistant à sélectionner un sous-ensemble d'un ensemble de données plus large. Le choix aléatoire du sous-ensemble ajoute de l'incertitude dans la mesure où un attaquant qui essaie d'identifier un individu ne pourra pas être sûr de la présence de l'individu[106]. Comme vu précédemment dans les types d'attaquant, la probabilité d'identification est beaucoup plus grand dans une attaque de type journaliste que dans une attaque de type procureur.
Les méthodes pour sélectionner le sous-ensemble sont diverses et variées. Par exemple, un algorithme peut être utilisé pour générer des numéros aléatoires et ces numéros serviront pour sélectionner les enregistrements[107].
Outre la diminution du risque d'identification, l'échantillonnage présente d'autres avantages. Par exemple, dans le cadre de sondages à adresser à la population, au lieu de solliciter tous les habitants d'un pays ou d'une région (ce qui comporte des coûts en termes d'argent et de temps), il est possible de recourir à un échantillon représentatif. C'est d'ailleurs ce qui a été mis en

[103] PRASSER/ KOHLMAYER/ A. KUHN, p. 3.
[104] KNIOLA, p. 1.
[105] PRASSER/ KOHLMAYER/ A. KUHN, p. 3.
[106] ISO/IEC 20889:2018, p. 4.
[107] *Ibid.*

place par l'OFS depuis 2010 : un cadre d'échantillonnage basé sur les données des registres des habitants des communes et cantons[108].
Le résultat de cette méthode se présente sous la forme de micro-données[109].

2. Agrégation

L'agrégation est une méthode qui consiste à combiner des attributs afin de fournir des informations qui sont moins détaillées de ce qui est effectivement observé. L'agrégation peut avoir lieu au niveau des valeurs d'un même type d'attribut ou bien au sein d'un même enregistrement en combinant divers attributs[110]. Un exemple d'agrégation au niveau des valeurs d'un même type d'attribut est le fait de calculer la moyenne de tous les salaires au sein d'un village. L'avantage en termes de confidentialité est de ne rien dévoiler sur un enregistrement en particulier. Un exemple d'agrégation combinant divers attributs d'un même enregistrement est ce qui a été fait dans l'affaire de la banque suisse qui voulait transférer des données aux autorités américaines, où les transactions relatives à un même compte ont été agrégées sur un mois.

a) Cas pratique d'agrégation dans le contexte de l'enregistrement des maladies oncologiques en Suisse

La loi fédérale du 18 mars 2016 sur l'enregistrement des maladies oncologiques (LEMO) ainsi que son ordonnance (OEMO) régissent l'enregistrement des maladies oncologiques en Suisse. L'article 30 de l'ordonnance spécifie les conditions d'anonymisation. Tout d'abord, d'après les al. 1 et 2, les identifiants directs ou indirects doivent être supprimés, à l'exception du mois et de l'année de naissance et de décès. En outre, d'après les al. 3 et 4, une agrégation doit être opérée en cas de communication à des tiers. Pour que des données agrégées soient considérées comme anonymes, elles doivent être constituées d'au moins 20 individus[111].

## D. Techniques de suppression

1. Masquage

Il s'agit d'une technique qui consiste à enlever tous les identifiants directs ou indirects d'un jeu de données[112]. Comme déjà mentionné dans l'introduction, cette technique à elle seule peut ne pas suffire pour protéger contre une re-identification lorsqu'elle se limite à effacer les identifiants directs et il faudra donc la combiner avec d'autres mesures.

a) Suppression locale

Ici la suppression se fait sur des valeurs spécifiques dans certains enregistrements précis. Généralement on applique cette technique aux valeurs rares ou aux combinaisons de valeurs rares. Il est également possible de supprimer l'ensemble de l'enregistrement [113].

2. Le cas de la réidentification du Gouverneur William Weld

En 1997, Latanya Sweeney, alors étudiante diplômée au Massachusetts Institute of Technology (MIT), a réussi à réidentifier des données médicales soi-disant anonymisées appartenant au Gouverneur William Weld. Sweeney a obtenu un ensemble de données de santé provenant de

[108] https://www.bfs.admin.ch/bfs/fr/home/bases-statistiques/recensement-population/recensement-element-systeme-vaste/registre-echantillonnage.html#:~:text=Depuis%202010%2C%20l'OFS%20dispose,mises%20%C3%A0%20jour%20chaque%20trimestre (consulté le 12 mai 2024).

[109] ISO/IEC 20889:2018, p. 5.

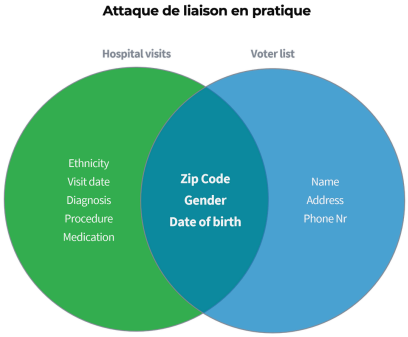
[110] *Ibid.*

[111] OFSP Questionnaire, Ordonnance sur l'enregistrement des maladies oncologiques (OEMO), RS 818.331.

[112] ISO/IEC 20889:2018, p. 8.

[113] *Ibid,* p. 15.

la Commission d'Assurance du Groupe du Massachussetts (GIC). Tous les identifiants directs comme les noms ou les adresses avaient était retirés. Les données devaient servir à la recherche pour améliorer les soins de santé ainsi que le contrôle des coûts. Suite à cela, elle a acheté une liste d'électeurs de Cambridge pour vingt dollars. Cette liste contenait des informations telles que le nom, l'adresse, le code postal, la date de naissance et le sexe de chaque électeur inscrit. En croisant ces deux ensembles de données, plus spécifiquement en cherchant les combinaisons entre code postal, date de naissance et sexe, elle a pu isoler les profils correspondants (comme illustré dans l'image suivante). Cette réidentification a pu démontrer que la combinaison de quasi-identifiants peuvent avoir un fort pouvoir identifiant, selon le contexte[114].




[115]

Il est important de noter cependant que Sweeney disposait d'une information cruciale qui a permis cette démarche : elle savait que le gouverneur avait été hospitalisé ainsi que le lieu et la date de l'hospitalisation. Elle pouvait donc raisonnablement supposer que le gouverneur se trouvait dans l'ensemble de données du GIC[116]. Pour reprendre les modèles d'attaquants, elle a effectué une attaque de type procureur qui a plus de chance d'aboutir en une réidentification qu'une attaque de type marketeur ou journaliste.
L'incident a influencé directement la réforme de la loi HIPAA (*Health Insurance Portability and Accountability Act*) des Etats-Unis qui a été modifiée pour y inclure des normes plus strictes d'anonymisation des données de santé[117].

### 3. Le cas de l'ensemble de données du prix Netflix

La plateforme Netflix a publié un ensemble de données anonymisé de plus de 100 millions de notes de films données par environ 500 000 abonnés dans le cadre d'un concours, le « prix Netflix ». Cette base de données devait permettre aux participants du concours de développer un algorithme de recommandations pour la plateforme. Un prix d'un million de dollars était ensuite versé à celui qui arriverait à présenter la meilleure proposition d'algorithme. Malgré l'anonymisation de la base de données, des chercheurs ont réussi à désanonymiser les données grâce à des informations externes[118].
Les identifiants directs tel que le nom avaient été supprimés de la base de données Netlfix. Chaque enregistrement contenait uniquement : le titre des films notés, les notes attribuées et la date de soumission des notes. D'ailleurs, un léger bruit avait été ajouté aux informations restantes[119].

---

[114] BARTH-JONES, p. 2ss.

[115] Prise sur le site https://www.syntho.ai/fr/5-examples-why-removing-names-is-not-an-option/ (consulté le 7 juin 2024).

[116] *Ibid*., p. 15.

[117] *Ibid.,* p. 1.

[118] NARAYANAN/ SHMATIKOV, p. 10ss.

[119] *Ibid.*

Des chercheurs ont recouru à des connaissances auxiliaires qui étaient publiquement accessible, notamment la plateforme Internet *Movie Database* (IMDb), une plateforme où les utilisateurs peuvent publier leurs avis et notes de films. Ces avis ne sont généralement pas anonymisés[120].
Ils ont ensuite développé un algorithme capable de corréler les notes publiques avec les notes anonymisées de Netflix. Un score de similarité était alors attribué à chaque paire d'enregistrements entre les deux bases de données. Les algorithmes ont été développés pour tolérer des erreurs. Par exemple, s'agissant de la date de soumission des notes, une marge de 14 jours était tolérée. La même chose valait pour les notes attribuées, ou une différence de 1 point était admis[121].
Il a alors été possible, avec seulement 6 à 8 films, d'identifier correctement jusqu'à 99% des enregistrements de la base de données Netflix[122].
Les chercheurs se sont ensuite posé la question suivante : pourquoi un utilisateur qui évalue des films sur IMDb, souvent sous son propre nom et prénom, devrait-il se préoccuper de la confidentialité des évaluations faites sur Netflix ? Ils ont alors étudié un individu en particulier et se sont aperçus que, en utilisant la base de données Netflix, ils pouvaient déterminer son orientation politique à partir d'opinions exprimées sur deux films, ainsi que son orientation religieuse à partir de deux évaluations de deux autres films. Ces conclusions n'auraient pas pu être tirées à partir de la base de données d'IMDb, car l'utilisateur n'avait pas évalué ces mêmes films sur cette autre plateforme[123].

## E. Généralisation

La généralisation est un concept regroupant plusieurs techniques et qui se base sur l'idée de remplacer les valeurs stockées par des alternatives sémantiquement cohérentes et véridiques, mais moins précises[124]. En d'autres termes, il y a une dilution de l'information parce que l'on modifie l'échelle ou l'ordre de grandeur d'une information. Cela constitue une bonne réponse au risque d'individualisation mais devra avoir une approche quantitative spécifique afin de prévenir le risque de corrélation et inférence[125].
Pour des données numériques, on pourrait par exemple fixer une valeur maximale au-delà de laquelle l'information n'est plus donnée spécifiquement. Par exemple : au-delà des salaires supérieurs à 10 000 CHF, il sera indiqué « supérieur à 10 000 CHF ». Il est également possible de généraliser des données non-numériques. Par exemple, au lieu d'indiquer la ville de naissance, on indique le pays[126].

## F. Randomisation

La randomisation constitue une autre catégorie de techniques d'anonymisation. Elle repose sur l'idée d'altérer la véracité des informations dans le but d'affaiblir le lien entre celles-ci est l'individu. C'est une solution efficace pour parer au risque d'inférence, cependant chaque enregistrement gardera sa propre singularité, ce qui constitue toujours un risque du point de vue de l'individualisation et de la corrélation. C'est pourquoi il peut se révéler important de combiner plusieurs techniques d'anonymisation.[127]

---

[120] *Ibid,* p. 15.
[121] *Ibid*, p. 1.
[122] *Ibid*, p. 12.
[123] *Ibid*, p. 16.
[124] PIERANGELA/SWEENEY, p. 2.
[125] WP 216, p. 4.
[126] ISO/IEC 20889:2018, p. 18.
[127] WP 216, p. 13.

D'ailleurs, dans la mesure où cette technique altère la véracité des attributs visés, elle doit être spécifiquement adaptée aux objectifs du cas pour pouvoir produire un jeu de données qui soit utile[128].

1. Ajout de bruit

Parmi les techniques de randomisation, il y a l'ajout de bruit qui consiste à introduire du bruit aléatoire dans les valeurs originales, sans toutefois compromettre les propriétés globales de la base de données. Un exemple est l'arrondi base-x[129], qui ajoute un bruit aléatoire en arrondissant les valeurs au multiple x le plus proche. Dans le tableau 2, l'âge réel a été arrondi en base-3[130].

| | | | |
|---|---|---|---|
| **1.** | 22 | → | 21 |
| **2.** | 41 | → | 42 |
| **3.** | 17 | → | 18 |

**Tableau 2**

Il faut faire attention à ajouter un bruit qui soit cohérent avec la logique des attributs. Par exemple, dans le cas précèdent, si nous avions décidé d'ajouter une valeur de 100 à tous les enregistrements, nous aurions obtenu des valeurs incohérentes. Ce qui représente un risque, car un attaquant pourrait s'en apercevoir et même être capable de filtrer le bruit[131].

2. Permutation

Dans la permutation, les valeurs des attributs sont mélangées entre différents enregistrements de sorte à avoir des valeurs qui sont artificiellement liées à un individu. Cette technique permet de garantir que la distribution des valeurs reste inchangée. Cela peut être intéressant dans la mesure où l'ajout de bruit aléatoire cohérent peut se révéler complexe[132]. Il faut cependant prêter attention à garder la relation logique entre deux attributs au sein du même enregistrement. Par exemple, l'âge et le nombre d'années d'expérience professionnelle sont logiquement liés. Une permutation mal effectuée pourrait communiquer des informations à l'attaquant qui pourrait être en mesure d'inverser la permutation[133]. Pour résoudre ce problème, dans l'exemple à peine illustré, il faudrait permuter le nombre d'années d'expérience professionnelle uniquement au sein d'un groupe appartenant à une même tranche d'âge[134].

## G. Modèle formel de mesure de la protection de la vie privée

1. Introduction

Un modèle formel de mesure de la protection de la vie privée est une méthode qui permet de calculer le risque de reidentification et dans certains cas, permet de fournir des garanties mathématiques contre la réidentification[135].

2. K-anonymat

Le k-anonymat vise à rendre impossible l'identification d'un individu dans un ensemble de données, même lorsque ces données sont croisées avec des sources d'information externes. Pour ce faire, il s'assure que chaque individu partage avec au moins k-1 autres individus chaque combinaison d'identifiants[136]. Pour obtenir cela, les attributs sont généralisés jusqu'à obtenir

---

[128] ISO/IEC 20889:2018, p. 18.

[129] La fonction suivante a été utilisée : **Arrondi_Base_x(n,x)=round(x/n)×x**, où «n» correspond au nombre à arrondir ; « x » la base à laquelle il faut arrondir et « round » la fonction d'arrondi qui arrondi au nombre entier le plus proche.

[130] KLEINER/ HEERS, p. 13.

[131] WP 216, p. 14.

[132] WP 216, p. 15.

[133] *Ibid.*, KLEINER/ HEERS, p. 13.

[134] KLEINER/ HEERS, p. 13.

[135] ISO/IEC 20889:2018, p. 20.

[136] SAMARATI/ SWEENEY, p. 5.

k-1 individus partageant la même valeur[137]. A noter que d'autres techniques d'anonymisation peuvent être utilisées pour atteindre une valeur k-anonyme[138].
Pour mieux comprendre ce modèle, nous allons l'illustrer par un exemple simple. Le tableau 3 correspond à une base de données médicale. Puisque les dates de naissance sont différentes, il est assez facile d'individualiser chaque enregistrement.

| Date de naissance | Sexe | Diagnostic |
|---|---|---|
| 01.03.1996 | M | Crise cardiaque |
| 02.04.1997 | M | Diabète |
| 03.03.1980 | M | Crise cardiaque |
| 24.02.1979 | M | Crise cardiaque |

**Tableau 3**

La tableau 4 correspond au tableau 3 après avoir été rendue 2-anonyme en généralisant la date de naissance à l'année de naissance.

| Année de naissance | Sexe | Diagnostic |
|---|---|---|
| 1990-2000 | M | Crise cardiaque |
| 1990-2000 | M | Diabète |
| 1970-1980 | M | Crise cardiaque |
| 1970-1980 | M | Crise cardiaque |

**Tableau 4**

Cette technique constitue une bonne solution contre l'individualisation dans la mesure où désormais il y a k utilisateurs qui partagent les mêmes attributs et par conséquent il n'est plus possible d'isoler un individu. Cependant, une base de données k-anonyme peut toujours être vulnérable à la corrélation et l'inférence. Par exemple, dans le tableau 2, si l'attaquant sait qu'un individu figure dans l'ensemble de données et est né en 1979, nonobstant le fait que l'année de naissance ait été généralisée, il saura que l'individu a fait une crise cardiaque.

3. L-diversité

La l-diversité étend le concept de k-anonymat afin de répondre au problème rencontré dans le dernier exemple. L'idée est d'avoir pour chaque groupe de k-anonymat au moins *l* valeurs distinctes pour les valeurs sensibles (dans notre exemple, le diagnostic)[139].
Par exemple, dans le tableau 4, le deuxième groupe de k-anonymat (les individus nés entre 1970 et 1980) a un l-diversité de 1, ce qui permet une attaque par inférence avec une certitude de 100%. Il suffit de savoir qu'un individu est né en 1979 et qu'il se trouve dans la base de données pour déterminer qu'il a fait une crise cardiaque. Afin d'avoir un l-diversité de deux, il faudrait avoir deux diagnostics différents.
La l-diversité vise à insérer un degré d'incertitude aux attaques par inférence, cependant des inférences probabilistes restent possibles[140]. Par exemple, toujours dans le tableau 4, si l'attaquant sait qu'un individu né en 1996 est dans la base de données, il peut établir avec une probabilité de 50% que cet individu a eu une crise cardiaque.

4. T-proximité

La t-proximité étend ultérieurement le concept de k-anonymat et l-diversité afin de répondre à certaines faiblesses de ces modèles. L'idée est qu'au sein d'une classe d'équivalence, la distribution des valeurs d'un attribut sensible ne diffère pas trop de la distribution globale de cet attribut dans la base de données[141].

[137] WP 216, p. 18.
[138] ISO/IEC 20889:2018, p. 20.
[139] WP 216, p. 18.
[140] *Ibid*.
[141] *Ibid*.

La t-proximité rajoute donc une contrainte supplémentaire : chaque classe d'équivalence ne doit pas uniquement comporter *l* valeurs différentes, mais chaque valeur doit être représentée autant de fois que nécessaire pour refléter la distribution originale. Le modèle est utile lorsque le base de données anonymisée doit garder les mêmes propriétés que l'originale[142].
La capacité de ce modèle de protéger contre les inférences est cependant limitée lorsque les valeurs des données sont inégalement distribuées. Par exemple, si des valeurs sont très fréquentes alors que d'autres sont très rares, un attaquant peut toujours en tirer des conclusions sur un individu[143].

5. Confidentialité différentielle

Jusqu'à présent nous avons vu des modèles où la mesure de protection de la vie privée s'effectue *avant* le partage/la publication du jeu de données. Il s'agit de modèles non-interactifs de mécanisme de protection de la vie privée. Ces modèles présentent cependant plusieurs faiblesses. Parmi ces faiblesses, il y a la difficulté de déterminer une utilité qui n'a pas encore été exprimée au moment de l'application des mesures de désidentification[144].
Dans le modèle interactif en revanche, une interface permet aux utilisateurs de poser des questions au détenteur sur les données afin d'obtenir des réponses qui seront brouillées en fonction de la sensibilité de la requête. Parmi ces modèles, il y a la confidentialité différentielle. Introduit en 2006 par quatre chercheurs dont deux appartenant à Microsoft, il s'agit d'un modèle mathématique qui vise à calculer le bruit qu'il faut ajouter à chaque requête afin de ne pas dépasser un certain seuil (appelé paramètre) de perte de confidentialité (qui est défini par politique)[145]. La promesse de la confidentialité différentielle est de garantir à la personne concernée qu'elle ne sera pas affectée par la participation à une étude ou une analyse et ce, indépendamment des autres données ou sources disponibles[146].
Le terme « perte de confidentialité » ne se réfère pas à une perte effective mais plutôt à une réduction de la probabilité que la confidentialité soit maintenue. L'accumulation de connaissance d'un hypothétique attaquant au fil du temps et des requêtes crée cette perte de confidentialité. Lorsque le « budget » de confidentialité (le paramètre) est atteint, la base de données devrait arrêter de répondre aux requêtes ou prendre d'autres mesures[147].
Les algorithmes de confidentialité différentielle fonctionnent en ajoutant une quantité de bruit aléatoire générée par une distribution de probabilité qui a été sélectionnée, afin de maintenir l'utilité des données. Il y a deux modèles envisageables[148] :

- **Modèle-serveur :** les individus soumettent leurs données personnelles à un « curateur » qui les stocke dans un serveur central. Une analyste va alors former des requêtes à l'adresse du curateur et ce dernier ajoutera du bruit avant d'envoyer les réponses. L'avantage de ce modèle est qu'il permet à l'algorithme de calculer la plus petite quantité de bruit. La concentration des données personnelles dans un seul endroit crée en revanche des soucis de sécurité. En outre le curateur n'est pas toujours considéré digne de confiance. Les individus peuvent être réticents à confier leurs données personnelles au curateur[149] ;

---

[142] WP 216, p. 18.

[143] ISO/IEC 20889:2018, p. 21.

[144] DWORK / MCSHERRY / NISSIM / SMITH, p.3ss.

[145] *Ibid.*

[146] DWORK/ ROTH, p. 20.

[147] ISO/IEC 20889:2018 p. 21.

[148] *Ibid.*

[149] https://www.nist.gov/blogs/cybersecurity-insights/threat-models-differential-privacy (consulté le 30 mai 2024)

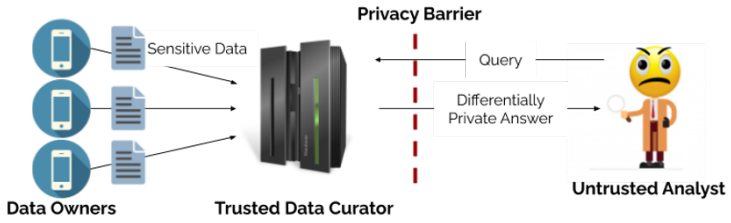



*Image représentant le modèle-serveur, reprise du National Institue of Standards and Technology* [150]

- **Modèle local** : ce modèle répond aux deux désavantages du modèle-serveur dans la mesure où le bruit est ajouté aux points de sortie des individus qui envoient leurs propres données. Cela signifie que le curateur recevra des données déjà bruitées et il ne sera plus nécessaire qu'il soit digne de confiance. D'ailleurs, si le serveur subit une violation de sécurité comme par exemple un accès illégal, les données visionnées seront bruitées. Le désavantage de ce modèle par rapport au modèle serveur est que le bruit ajouté aux données est beaucoup plus grand[151].

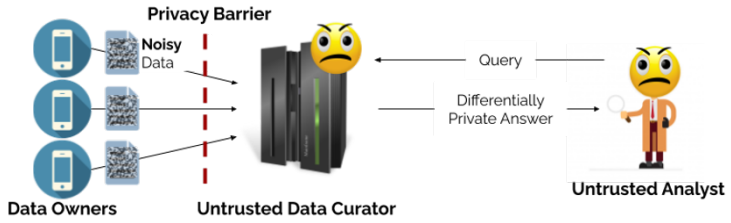



*Image représentant le modèl local, reprise du National Institue of Standards and Technology* [152]

## H. Pseudonymisation

Le terme « pseudonymisation » se réfère à une catégorie de méthodes[153] qui consistent à remplacer des identifiants par un code (pseudonyme), afin d'empêcher l'identification[154]. Généralement c'est un attribut unique, typiquement le nom, qui est remplacé[155]. Simultanément, une table de correspondance est créée afin de lier le nom avec le pseudonyme. Ainsi, seules les personnes qui ont accès à la table pourront renverser la pseudonymisation[156]. Cette technique permet de corréler des enregistrements provenant de la même personne dans différentes bases de données sans révéler son identité[157].

La pseudonymisation crée de l'information supplémentaire : la table de correspondance ou la clé cryptographique. L'accès à ces informations doit être protégé par des mesures techniques et organisationnelles adéquates[158].

Dans la mesure où la personne est toujours susceptible d'être réidentifiée, la pseudonymisation diffère de l'anonymisation[159]. Cependant, la pseudonymisation réduit le risque de mise en corrélation d'un ensemble de données avec l'identité originale de la personne concernée. Il

---

[150] *Ibid.*

[151] *Ibid.*

[152] *Ibid.*

[153] ISO/IEC 20889:2018 p. 15.

[154] CR LPD – MEIER/TSCHUMY, art. 5 N 29.

[155] *Ibid* ; WP 216, p. 22.

[156] PFPDT MTO 2024 p. 25ss.

[157] ISO/IEC 20889:2018 p. 15.

[158] *Ibid.*

[159] AEPD *10 Misunderstandings,* p. 3.

s'agit donc d'une bonne mesure de sécurité[160]. Plusieurs méthodes sont envisageables pour établir la correspondance entre les données identifiantes et leurs pseudonymes[161]. Le pseudonyme peut être totalement indépendant de l'attribut qu'il veut remplacer. C'est le cas lorsqu'on recourt à un compteur où à générateur de nombres aléatoires. Dans ce cas une table de correspondance doit être créée. Le pseudonyme peut dériver de l'attribut : c'est le cas lorsqu'on recours à des techniques de cryptographie comme une fonction de hachage ou un algorithme de chiffrement. Lorsqu'on recourt à un algorithme de chiffrement, la clé devra être gardée en sécurité[162].
Comme vu dans l'introduction, certaines fonctions de hachage ne résistent pas à des attaques par dictionnaire.

## I. Conclusion intermédiaire

Les techniques de désidentification sont nombreuses et variées. Lorsqu'on s'est aperçu que la simple suppression d'identifiants directs n'était plus suffisante à garantir l'anonymat, d'autres techniques pour désanonymiser des ensembles de données ont commencé à se développer. D'ailleurs, il s'agit d'un domaine de recherche qui est très actif, tout comme son pendant, la recherche sur les méthodes de réidentification. Ce qui implique qu'une base de données bien anonymisée aujourd'hui pourrait ne plus l'être demain. Les responsables du traitement doivent donc réévaluer régulièrement les risques associés à une base de données anonymisée.
Ce qui ressort de l'étude de cas vu précédemment est qu'un processus d'anonymisation doit être pensé selon le cas d'espèce. Il doit être taillé sur mesure en prenant en compte différents facteurs (utilité recherchée, attaquants probables ainsi que leurs connaissances auxiliaires). Lorsqu'une technique donnant comme résultat des macro-données est possible, elle devra être préférée, puisqu'elle assure plus de protection.
Parmi les facteurs à prendre en compte pour choisir les méthodes d'anonymisation appropriées, il y a les attaquants ainsi que leurs connaissances auxiliaires. Ces potentiels attaquants dépendront fortement du contexte de divulgation des données, comme on va le voir dans le prochain chapitre.

[160] WP 216, p. 20.

[161] PFPDT MTO 2020 p. 25ss.

[162] ENISA *pseudonymisation,* p. 22.

# IV. Différents degrés de divulgation d'un jeu de données

## A. Introduction

L'anonymisation d'un jeu de données doit être précédé d'une phase conceptuelle où plusieurs éléments sont étudiés afin de déterminer la méthode la plus adaptée. Parmi les éléments à prendre en compte, il y a l'environnement dans lequel les données vont être utilisées : resteront-elles au sein de l'entreprise/organisation, seront-elles divulguées à un ou plusieurs acteurs spécifiques ou seront-elles publiées sur Internet ? Prendre en compte l'environnement est essentiel dans la mesure où les connaissances et compétences d'un attaquant dépendront de ce contexte (par exemple, lorsque des données sont rendues publiques, on présumera de plus grandes connaissances et compétences compte tenu de la quantité et diversité d'acteurs potentiels)[163].

D'ailleurs, le choix de l'environnement dépendra de l'objectif recherché par le partage de données. En effet, comme mentionné précédemment, il existe une tension entre utilité des données et risque pour la confidentialité lorsqu'on utilise des techniques d'anonymisation. Selon l'objectif visé, il sera plus ou moins important de disposer de données suffisamment détaillées. Le responsable du traitement devra donc évaluer quelle option de divulgation est la plus appropriée[164]. A une extrémité du spectre, des données pseudonymisées pourraient être utiles pour des chercheurs qui recherchent une certaine granularité dans les données. Cependant, le risque d'identification est élevé. A l'autre extrémité du spectre, les données agrégées présentent un risque relativement faible (selon la taille de l'échantillon). Ces données seront alors relativement sûres mais pourraient ne pas offrir le niveau d'utilité recherché[165].

Nous allons aborder les différents types de divulgation possibles ainsi que le relatif risque de réidentification. Sans prétention d'exhaustivité, vu le vaste nombre de possibilités de divulgation, nous nous concentrerons sur une sélection trouvée dans la littérature.

## B. Données ouvertes (*open data*)

Une définition largement répandue des données ouvertes est celle de la *Open Knowledge Foundation*[166], également reprise par le Conseil fédéral dans sa Stratégie *Open Government Data*. En substance, le terme de données ouvertes (*Open data*) désigne les données qui peuvent être utilisées, éditées, analysées et transmises librement, sans aucune restriction que ce soit légales, financières ou techniques. Du point de vue légal, les données doivent être publiées sous licence libre. Du point de vue technique, il faut qu'elles puissent être traitées par ordinateur. Du point de vue financier, elles doivent être gratuites[167].

Les données ouvertes peuvent se révéler utiles dans de nombreux secteurs. Par exemple, publier des données gouvernementales en *Open access* peut améliorer considérablement la transparence de l'État et permettre aux citoyens de mieux comprendre les actions et décisions des autorités[168].

Les données ouvertes peuvent en outre constituer un catalyseur d'innovation car elles fournissent des ressources essentielles pour les entreprises ainsi que les chercheurs[169]. Elles permettent également d'améliorer l'offre de services déjà existants[170].

---

[163] Stiftung Datenschutz *Basic rules*, p. 6

[164] ICO *Code of practice,* p. 36.

[165] *Ibid.*

[166] https://okfn.org/en/library/what-is-open/ (consulté le 31 mai 2024)

[167] FF 2019 855.

[168] LAKOMA/ KALLBERG, p. 1.

[169] *Ibid.,* p. 3.

[170] PAPE/ SERNA-OLVERA/B. TESFAY, p. 1.

Cependant, les données ouvertes soulèvent également un grand nombre de défis, notamment :

- La qualité des données ne peut pas toujours être garantie ;
- Les formats dans lesquels les données sont stockées sont souvent obsolètes ;
- L'identification des sources n'est pas toujours possible ;
- La diversité des lieux de stockages rend leur accès difficile [171].

Parmi ces défis, il y a également des défis juridiques comme la question de la protection de la vie privée[172], la compatibilité avec les principes de la LPD dont celui de finalité et reconnaissabilité pour la personne concernée[173] ainsi que le respect du devoir d'information de l'art. 19 LPD. D'après le principe de **finalité**, les données ne peuvent être collectées et traitées que dans le but qui est indiqué lors de la collecte ou qui est prévu par la loi ou qui ressort des circonstances (détermination du but). Il est en outre interdit de modifier ce but en cours de traitement (immutabilité du but). Tandis que le principe de **reconnaissabilité** implique que la collecte ainsi que les finalités du traitement soient *reconnaissables* pour la personne concernée. Ce principe peut être satisfait lorsque la reconnaissabilité découle des circonstances, mais lorsque ce n'est pas le cas, il faudra prévoir une information active (par exemple à travers une politique de confidentialité). Le **devoir d'information** implique que les personnes concernées soient informées de l'identité du responsable du traitement, la finalité du traitement ainsi que les destinataires auxquels les données seront transmises (art. 19 LPD).

Or, lorsqu'on publie des données en libre accès, la *multitude d'utilisations possible* fait que la finalité du traitement ne peut être déterminée à l'avance[174]. Le devoir d'information se heurte à un autre obstacle qui est que l'on ne sait pas quel tiers va traiter des données et dans quel but. Est-ce qu'une information de type « catégorie de destinataire =potentiellement tous » et « but du traitement= tout type de traitement ») est envisageable ? Si l'on applique par analogie le raisonnement de MEIER sur le devoir d'information dans le contexte du Big Data, le caractère général et abstrait de ce type d'information demeure problématique[175].

L'on constate alors une tension entre la protection des données et l'*Open data*. En règle générale, les données ouvertes doivent donc être limitées aux données factuelles[176] où être anonymisées de sorte à pouvoir sortir du champ d'application de la LPD. Des techniques robustes d'anonymisation devront être mises en place[177]. En règle générale, les données pourront être publiées sous forme agrégée. En revanche, des données plus granulaires ne semblent pour l'instant pas être compatibles avec l'*Open data*[178]. Déterminer le niveau d'agrégation de données ou quelle autre technique choisir peut se révéler une tâche complexe[179]. Afin de s'orienter dans cette démarche, certains organismes ont développé des outils pour évaluer le risque d'identification dans un jeu de données[180].

A noter qu'au vu des avancées technologiques, l'évaluation doit être revue périodiquement[181]. Dans le cas d'un jeu de données qui devait se révéler compromis, il serait nécessaire de le retirer

---

[171] https://bigdata-dialog.ch/fr/les-defis-des-donnees-de-recherche-ouvertes/ (consulté le 31 mai 2024)

[172] *Ibid.*

[173] ERARD / HEUSGHEM / PARISATO, p. 5.

[174] *Ibid.*, WP 203, p. 48.

[175] MEIER (Big Data), p. 65.

[176] https://www.edoeb.admin.ch/edoeb/fr/home/kurzmeldungen/nsb_mm.msg-id-86956.html (consulté le 20.05.2024)

[177] ZEVENBERGEN/ BROWN/ WRIGHT/ ERDOS, p. 15.

[178] WP 203, p. 49.

[179] WP 207, p. 12.

[180] Voir par exemple l'outil développé par SPHN : https://sphn.ch/document/template-use-case-evaluation-and-risk-assessment/ (consulté le 7 juin 2024).

[181] WP 203, p. 49.

de la plateforme ainsi que de communiquer aux utilisateurs d'arrêter le traitement. C'est une étape nécessaire pour limiter les dommages même si la faisabilité ainsi que l'efficacité de cette démarche une fois que « *le mal est fait* » semblent limitées[182].
Nous pouvons également nous poser la question du consentement comme motif justificatif. Dans la protection des données en effet, ils existent des motifs justificatifs qui peuvent venir réparer le non-respect des principes généraux tel que la finalité ou la reconnaissabilité et rendre par conséquent le traitement licite. Peut-on consentir à mettre nos données personnelles en *Open Data* ? Selon la doctrine traditionnelle (également appliquée sous l'angle de l'art. 28 al. 2 CC), le consentement doit porter sur un objet déterminé, être suffisamment prévisible, être clair et non équivoque. D'ailleurs, plus l'atteinte est grave et plus le consentement doit être déterminé quant à son objet[183]. A l'heure actuelle, il me semble donc difficile de concevoir un consentement valide pour traiter des données personnelles en *Open Data*, qui par sa définition est un environnement très vaste et presque indéfini, surtout lorsqu'il s'agit de données sensibles qui ont un potentiel d'atteinte à la personnalité plus grand.

1. Exemple du libre accès aux données publiques en Suisse (projet *Open Government Data*)

Le Conseil fédéral a jusqu'alors adopté deux « Stratégies de libre accès aux données publiques suisses » pour les périodes 2014-2018 et 2019-2023[184] ainsi qu'un plan directeur 2024-2027[185]. L'objectif principal est de promouvoir sa politique en matière de libre accès aux données publiques. Pour ce faire, le Conseil fédéral a développé une infrastructure centralisée : le portail *opendata.swiss* qui répertorie (sans les stocker) toutes les données en libre accès de l'administration publique suisse[186].
Les données publiques sont considérées être librement accessibles lorsque deux conditions cumulatives sont remplies :

1) Leur accès est libre ;
2) Leur utilisation n'est pas limitée notamment pour des raisons relevant du droit de la protection des données, ce qui permet à des tiers de les réutiliser librement[187].

Le Conseil fédéral semble donc parfaitement conscient que la stratégie de libre accès doit être en parfaite conformité avec les prescriptions relatives à la protection des données. Il reconnait que les publications sont généralement faites sous forme agrégée et préalablement anonymisée. L'autorité qui publie doit alors prendre toute mesure technique et organisationnelle appropriée pour éviter une divulgation de données personnelles[188].
Les données disponibles sur *opendata.swiss* semblent donc se limiter à des données sur la topographie, la météorologie et la géolocalisation[189].
Constatant que *l'Open Government Data* était ralenti par l'incertitude juridique - grand nombre d'autorités ne publiaient pas toutes les données par crainte de publier des données personnelles[190]- la Loi fédérale du 17 mars 2023 sur l'utilisation des moyens électroniques pour l'exécution des tâches des autorités (LMETA) a été adoptée et devrait permettre de clarifier le cadre légal. En particulier, l'art. 10 LMETA instaure une obligation pour les unités

[182] WP 207, p. 18.
[183] MEIER (Big Data), p. 67.
[184] FF 2019 855, FF 2014 3347.
[185] OFS Masterplan 2024-2027.
[186] FF 2019 855.
[187] FF 2014 3347, p. 3.
[188] FF 2014 3347, p. 11.
[189] MARMIER/ METTLER, p. 9.
[190] *Ibid.*

administratives de publier les données qu'elles collectent ou produisent dans l'exécution de leurs tâches. Toutefois, cette obligation est exclue lorsqu'il s'agit de données personnelles (al. 2).

## C. Accès semi-public

Un jeu de données peut être publié de façon « semi-publique ». Cela signifie que les données sont mises à disposition pour téléchargement pour tout le monde, à condition de s'enregistrer auprès de la plateforme et d'accepter les restrictions concernant le traitement et le partage des données (généralement sous la forme de conditions d'utilisation)[191].
Des mesures supplémentaires de confidentialité et de sécurité peuvent être exigées dans les conditions d'utilisations, cependant elles semblent difficiles à mettre en œuvre et partant ce type d'ouverture offre une protection limitée[192]. Les conditions doivent à *minima* interdire le récipiendaire de :

- Tenter de réidentifier les individus dans l'ensemble de données ;
- Établir des liens avec des informations externes ;
- Partager l'ensemble de données sans autorisation[193].

Nonobstant les conditions d'utilisation, la probabilité de réidentification pour ce type de divulgation est considérée équivalente à celle d'un jeu de données en libre accès[194].

### 1. Exemple de la plateforme *swissubase.ch*

Swissubase est une plateforme dédiée à la gestion des données de recherche en Suisse. Crée par FORS, les Universités de Lausanne, Neuchâtel et Zürich, elle vise à faciliter le partage et la préservation des données de recherche pour une réutilisation facilité dans la communauté scientifique suisse[195].
Dans leur guide d'utilisation destiné aux personnes souhaitant déposer un jeu de données, la plateforme prend plusieurs mesures afin de protéger les données personnelles :

- Le dépositaire doit confirmer avant publication qu'il n'y aucune donnée personnelle ou que les personnes concernées ont été informées de la publication et ont donné leur accord ;
- Le dépositaire peut restreindre le droit de téléchargement des données à trois catégories : « aucune restriction » ; « recherche académique et enseignement » et « recherche académique ». Ces options offrent un contrôle supplémentaire sur qui peut accéder aux données ;
- Le dépositaire peut également choisir de permettre le téléchargement sous condition d'obtenir un accord préalable. S'il choisit cette option, il sera notifié d'une demande et pourra accepter ou non, ce qui renforce ultérieurement le contrôle[196].

Contrairement à la définition de données semi-publiques, l'accès aux données de *swissubase.ch* est donc plus limité. L'environnement du jeu de données est donc plus restreint et cela a pour conséquence de limiter le risque de réidentification.

## D. Partage limité avec des entités précises (partage « non public »)

Lorsqu'on parle de partage non-public, d'après le guide publié par le Commissaire à l'information et à la protection de la vie privée de l'Ontario, on se réfère à un partage qui est

---

[191] *IPC Ontario Guidelines,* p. 7.

[192] *Ibid.*

[193] *Id.,* p. 20.

[194] *Id.,* p. 15.

[195] https://resources.swissubase.ch/about-us/?lang=fr (consulté le 31 mai 2024)

[196] Swissubase *User Guide,* p. 9ss.

limité à des entités spécifiques. Dans ce contexte, il y a généralement un contrat qui régit l'accord de partage des données, ce qui crée une protection plus grande que dans les modèles vus précédemment[197]. Il s'agit d'un élément important afin de mitiger le risque de réidentification[198].

D'ailleurs, plus le contenu de l'accord de partage de données est strict en termes d'exigences de confidentialité, plus la probabilité qu'une tentative de réidentification soit lancé volontairement par le récipiendaire est faible. A titre d'exemple, les contrôles de sécurité et de confidentialité qui peuvent être inclus dans un accord sont :

- limitation d'accès du jeu de données au sein de l'organisation/entreprise récipiendaire limitée au personnel autorisé ;
- des accords de non-divulgation sont fait signer aux membres du personnel ayant accès au jeu de données ;
- la conservation du jeu de données est limitée dans le temps[199].

A noter que d'autres facteurs déterminent également le risque d'une tentative de réidentification par le récipiendaire : les capacités ainsi que la motivation de ce dernier[200].

D'ailleurs, si le récipiendaire devait ne pas respecter l'accord, le responsable devrait alors agir en exécution du contrat[201].

La relation entre le récipiendaire et celui qui partage les données peut être de nature très variée. A titre d'exemple, l'on peut partager des données :

- Avec un individu au sein de la même organisation/entreprise ;
- Avec un *constortium* de recherche ;
- Avec une entité commerciale ou gouvernementale[202].

Lorsque la relation est contractualisée, dans tous les cas, le niveau de sécurité ainsi que de contrôle sera plus élevé par rapport à un accès semi-ouvert, en raison de la capacité à imposer des obligations contractuelles strictes ainsi qu'à la possibilité de mettre en œuvre des mesures de sécurité personnalisées.

[197] *IPC Ontario Guidelines,* p. 7.

[198] *Ibid*., p. 11, ZEVENBERGEN/ BROWN/ WRIGHT/ ERDOS, p. 28.

[199] *IPC Ontario Guidelines,* p. 17.

[200] *Id*., p. 18.

[201] ZEVENBERGEN/ BROWN/ WRIGHT/ ERDOS, p. 32.

[202] *Id.,* 28.

# V. Conclusion

Comme nous l'avons déjà mentionné, la protection des données personnelles est un enjeu crucial dans notre société qui gravite autour de l'information. La notion de donnée personnelle est extrêmement large et englobe aussi bien les identifiants directs que les identifiants indirects. Le principal défi réside dans la notion d'identifiabilité de la donnée personnelle. A quel moment un ensemble d'identifiants indirects devient-il une donnée personnelle ? Trouver une réponse adaptée à cette question permet d'éviter que la loi sur la protection des données s'applique à l'infini, tout en garantissant qu'elle s'applique correctement là où cela est nécessaire pour protéger la personnalité des personnes concernées. Une approche contextuelle adaptée à chaque spécificité du cas semble alors la meilleure solution.

D'ailleurs, comme on l'a pu voir dans le dernier chapitre, le risque de réidentification dépend largement de l'environnement de divulgation. Plus les données sont largement partagées, moins on est capable de maîtriser cet environnement, ce qui crée un risque majeur de réidentification qui nécessitera des techniques d'anonymisation plus robustes ainsi que d'autres mesures de type organisationnel et légal (contrat, droits d'accès, …).

Par ce travail de Master, je voulais essayer de comprendre si le monde juridique était bien aligné avec le monde technique s'agissant de la notion d'identifiabilité. Quand bien même les tribunaux auront très certainement l'occasion de se reprononcer sur cette définition, il me semble qu'en adoptant une approche contextuelle, ils s'orientent vers une juste direction.