Anonymisation des données : enjeux d'éthique pour la recherche scientifique

Christine Froidevaux, Jean-Gabriel Ganascia, Claude Kirchner

Anonymisation des données : enjeux d'éthique pour la recherche scientifique

Christine Froidevaux, Université Paris-Saclay — LISN
christine.froidevaux@lisn.fr

Jean-Gabriel Ganascia, Sorbonne Université — LIP6
jean-gabriel.ganascia@lip6.fr

Claude Kirchner, Inria
claude.kirchner@inria.fr

31 août 2024

Membres du groupe de travail « anonymisation » de la CERNA (2018-2019) :
Christine Balagué, Danièle Bourcier, Max Dauchet, Gilles Dowek, Christine Froidevaux, Jean-Gabriel Ganascia, Claude Kirchner, Claire Levallois-Barth, Alice René, Félicien Vallet, Célia Zolynski.

Comment citer ce document :
Christine Froidevaux, Jean-Gabriel Ganascia, Claude Kirchner. *Anonymisation des données : enjeux d'éthique pour la recherche scientifique*. En collaboration avec Christine Balagué, Danièle Bourcier, Max Dauchet, Gilles Dowek, Claire Levallois-Barth, Alice René, Félicien Vallet, Célia Zolynski. Rapport technique. 31 août 2024.

Table des matières

Préambule historique

Les informations numérisées, les données, jouent un rôle fondamental dans notre civilisation numérique. Le respect de la conformité de leur utilisation aux régulations telles que le RGPD se pose chaque jour avec davantage d'acuité et l'anonymisation des données y joue un rôle central. C'est dans ce contexte qu'en 2018, la Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene (la CERNA, créée sous l'égide de l'alliance Allistene en 2012) s'est auto-saisie de l'étude des enjeux d'éthique de l'anonymisation des données, tout particulièrement dans le contexte de la recherche scientifique. Un groupe de travail a été mis en place par la CERNA en 2018 pour instruire le sujet et une journée d'étude « Anonymisation des données de recherche à caractère personnel » a été organisée par la CERNA le 3 juillet 2019 à l'Institut Telecom à Paris. Ce groupe de travail était composé de : Christine Balagué, Danièle Bourcier, Max Dauchet, Gilles Dowek, Christine Froidevaux, Jean-Gabriel Ganascia, Claude Kirchner, Claire Levallois-Barth, Alice René, Félicien Vallet et Célia Zolynski.

Suite à la mise en place, à la demande du premier ministre, en décembre 2019 du Comité National Pilote d'Éthique du Numérique (CNPEN), le comité de pilotage d'Allistene a décidé, sur proposition du président de la CERNA, de suspendre les travaux de la CERNA. Ce travail de réflexion sur l'anonymisation a été poursuivi et finalisé de 2020 à 2024 par les rapporteurs initiaux, Christine Froidevaux et Jean-Gabriel Ganascia, accompagnés de Claude Kirchner, indépendamment des travaux du CNPEN.

1 Introduction

Protection de l'intimité, de la vie privée, des données personnelles, des données sensibles, droit à l'image, droit à l'oubli par effacement des contenus, droit d'accès, droit de rectification, droit d'opposition, droit de limitation des traitements, droit à l'information en cas de faille de sécurité ou de violation de données etc. autant d'exigences légitimes qui, tout en suscitant l'adhésion, demandent à être mises en balance avec d'autres que beaucoup partagent aussi, comme l'aspiration à la transparence, à l'expression publique de soi ou à la sécurité. La sensibilité à toutes ces notions varie selon les cultures, les pays et les époques. Songeons que l'idée d'un droit à l'intimité (« Right to Privacy ») ne date que de 1890, avec l'article princeps de Warren and Brandeis [1] et que le droit à l'image n'existe pas vraiment dans les pays anglo-saxons. Qui plus est, avec le numérique et les évolutions contemporaines, ces notions prennent des connotations nouvelles qu'il convient d'élucider.

L'anonymat, c'est-à-dire l'absence de liens entre des informations et une personne, protège cette personne contre des agressions extérieures, et offre diverses possibilités attractives, jouer par exemple de façon anonyme ou écrire de manière anonyme ou sous pseudonyme selon le droit du code de la propriété intellectuelle relatif aux œuvres de l'esprit. Ce faisant, il permet aussi d'éviter de répondre en permanence de son comportement et par conséquent d'éluder sa responsabilité, ce qui peut conduire à toutes sortes de dérives. Cela peut aussi empêcher l'État d'assurer la sécurité des citoyens, sécurité qui relève d'une de ses prérogatives régaliennes majeures.

Au-delà de ces questions de principe, qui sont d'ordre social, politique et juridique, dans le monde numérique actuel, un nombre croissant d'obstacles techniques empêchent d'assurer l'anonymat, la confidentialité et le respect des données à caractère personnel.

Cela tient d'abord à ce qu'au plan scientifique, il existe des données, par exemple les données génomiques, les paramètres biométriques, les IRM du crâne [2] ou les images de visages, qui, par essence, ne peuvent pas être rendues anonymes. De même, l'internet des choses (aussi appelé internet des objets) [3] et les objets connectés rendent de plus en plus difficile l'effacement du lien entre les données générées et la personne qu'elles tracent au quotidien.

Cela tient ensuite à ce que les masses de données nécessaires à l'apprentissage machine reposent sur l'ouverture, le partage et la réutilisation de toutes les données, exigences peu compatibles avec les principes de finalité et de proportionnalité en vigueur dans la protection des données personnelles.

Cela tient enfin à ce que des données apparemment anonymes, c'est-à-dire sans référence explicite au nom d'une personne, comme par exemple des dossiers médicaux d'où le nom aurait été éliminé, permettent, par croisement avec d'autres informations, de retrouver l'identité des personnes. Dès lors, cette possibilité peut rendre vains les efforts faits pour garantir l'anonymat.

Les sciences du numérique jouent un rôle central tant dans la production et le traitement des données que dans leur anonymisation.

C'est dans ce contexte que la CERNA s'est autosaisie en 2018 des enjeux d'éthique de l'anonymisation

1. Samuel D. Warren et Louis D. Brandeis. « The right to privacy ». In : *Harvard Law Review* 4.5 (1890), p. 193-220. URL : http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm.

2. Christopher G. Schwarz et al. « Identification of Anonymous MRI Research Participants with Face-Recognition Software ». In : *New England Journal of Medicine* 381.17 (2019). PMID: 31644852, p. 1684-1686. URL : https://doi.org/10.1056/NEJMc1908881.

3. *Internet of Things* (IoT).

des données pour les scientifiques. Dans cet avis, nous nous attachons à :

- préciser les notions d'anonymat, de vie privée, de confidentialité et d'identités (identités plurielles, pseudonymes, avatars, etc.), aux plans sociologiques, normatifs et juridiques, tout en les plaçant dans une perspective historique, psychologique et culturelle ;
- indiquer les possibilités technologiques actuelles d'anonymisation, leurs évolutions et leurs conséquences en référence notamment :
  - au plan individuel (en précisant ce que l'on entend par anonymat, identités multiples et pseudonymes sur internet) ;
  - aux progrès en matière d'identification faciale et/ou vocale et à leurs extensions sur les réseaux sociaux qui intègrent des contenus multimédias ;
  - à tout ce qui concerne l'anonymisation et la pseudonymisation des données personnelles (en indiquant les techniques employées comme la k-anonymisation ou la *differential privacy*) et les techniques de ré-identification associées qui marquent leurs limites ;
  - à l'utilisation des profils comportementaux construits à partir de traces de navigation sur internet ou d'utilisation d'objets connectés.

Nous faisons des recommandations d'abord pour les scientifiques, mais aussi pour les citoyens et les institutions en vue d'une meilleure prise de conscience des enjeux de l'anonymisation des données. De plus, nous suggérons des pistes de recherche aux scientifiques. Elles sont explicitées au cours du texte et synthétisées à la fin du rapport.

2 De l'anonymat à l'anonymisation

2.1 Anonymat : repères historiques

Commençons par l'étymologie : attesté en français depuis le XVI^{e} siècle, le terme *anonyme* dérive du bas latin *anonymus* qui, lui-même, provient du grec *anônumos*, mot forgé à partir du préfixe privatif *an* et de *onoma*, « nom », ce qui signifie littéralement « sans nom ». Rien donc de vraiment nouveau derrière ce mot : **Est anonyme ce qui ne peut être assigné à un individu reconnaissable**.

Ainsi, parle-t-on depuis longtemps de **lettres anonymes** ou d'**œuvres anonymes** :

— dans un cas, quelqu'un se cache volontairement pour avancer masqué,

— dans l'autre, l'identité de l'auteur se perd pour différentes raisons.

Quoi qu'il advienne, l'idée demeure, et la connotation se révèle souvent négative lorsque cela revient à une *dissimulation* ou à une *perte*. Rappelons que, pendant longtemps, celui qui se dérobait délibérément aux regards apparaissait comme suspect : seul un coupable était censé procéder de la sorte. À tout moment, l'individu innocent devait agir au grand jour, en toute transparence, et rendre des comptes. À titre d'illustration, à Athènes, dans l'Antiquité, la plupart des votes se faisaient publiquement, à main levée, car les hommes libres devaient répondre de leurs choix.

Et, même dans le passé récent, l'anonymat n'avait qu'un caractère dérogatoire. On tolérait l'anonymat lors de **franchises temporaires**, à caractère plus ou moins ludique, comme les bals masqués. Jeu, écart passager, cela restait, bien évidemment, exceptionnel. On s'éloignait pour un temps de la morale ordinaire pour mieux y retourner après.

Il arrivait aussi que, dans de rares cas, l'anonymat ait des **connotations positives**, comme pour les dons qui, lorsqu'ils étaient anonymes, attestaient d'une sincère compassion, sans aspiration à une gratification ou à une quelconque reconnaissance.

Dans un autre registre, en France, depuis 1831, la **légion étrangère** engage ses hommes sous une simple « identité déclarée », sans exiger de preuves, ce qui, sous couvert d'anonymat, offre à certains une seconde chance. On peut aussi citer l'**accouchement sous X** qui permet à une mère de mettre au monde un enfant sans lui transmettre son nom, pour ne pas établir de lien officiel de filiation et procéder légalement à un abandon. Soulignons que la révision des lois de bioéthique de 2021 (LOI n° 2021-1017 du 2 août 2021 relative à la bioéthique) [1] ne permet plus le don de gamètes (ovule ou spermatozoïde) anonyme, tout en conservant à une femme la possibilité d'accoucher sous X et de conserver le secret de la filiation, si elle le souhaite.

2.2 Anonymat et numérique

Aujourd'hui, dans l'univers actuel, le **statut de l'anonymat change**. Sur internet le caractère ludique de l'anonymat semble demeurer. En témoigne l'usage généralisé des pseudonymes dans les jeux en ligne ou les réseaux sociaux. Il arrive même que certains changent d'âge, d'occupation ou de genre. Cependant, là où l'anonymat revêtait surtout une connotation négative, il prend désormais une connotation plutôt positive, notamment lorsqu'il existe un tiers de confiance qui connaît la personne, et s'engage à ne pas divulguer son identité (notion de **confidentialité**).

1. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043884384

Notons d'abord que de **nombreux dons s'accompagnent d'une volonté d'anonymat** ou plus exactement de **confidentialité**. Ainsi en va-t-il en France des dons :

— de gamètes, même si, au terme de la révision de la loi de bioéthique du 2 août 2021 [2]
« Les personnes qui souhaitent procéder à un don de gamètes ou proposer leur embryon à l'accueil consentent expressément et au préalable à la communication de ces données et de leur identité » et si « En cas de refus, ces personnes ne peuvent procéder à ce don ou proposer cet accueil. »
— d'organes et, dans une moindre mesure,
— de sang ;

qui se font uniquement en France sous couvert d'anonymat ou tout au moins de confidentialité.

Il est à noter que jusqu'à la plus récente révision de la loi de bioéthique du 2 août 2021 [3], le don de gamètes était confidentiel en ce sens que les enfants conçus par assistance médicale à la procréation n'accédaient pas automatiquement à l'identité des donneurs, même s'ils en faisaient la demande.

De même, souvent, l'**expertise** de dossiers ou de projets n'apparaît non biaisée que lorsqu'elle est confidentielle et s'accompagne de garanties d'anonymat pour les évaluateurs. Songeons à **l'évaluation scientifique** par des pairs des articles ou des projets. Cela part du principe que seul celui qui se prononce à l'abri des regards est libre : sinon, il pourrait être exposé à des pressions ou des représailles.

En politique, on pense aussi que l'anonymat est nécessaire pour pouvoir agir de façon désintéressée, au regard de ses seules convictions, sans subir de pression, ni *a fortiori* de rétorsions.

Ainsi, de nos jours, un vote n'apparaît démocratique que si on peut assurer que nul n'est en mesure de remonter à l'identité des personnes qui ont voté de telle ou telle façon. Lorsqu'on cherche à mettre en place un scrutin numérique (ou vote électronique), on se heurte à de nombreuses difficultés que des sénateurs ont décrites dans un rapport intitulé "*L'épineux dossier des machines à voter*" [4] : en effet, il faut s'assurer que la machine comptabilise correctement le suffrage émis par chacun des votants, sans pour autant autoriser quiconque à accéder à ces suffrages.

Dans ce contexte, il est révélateur qu'un groupe de hackers subversifs procédant à des attaques informatiques sur les réseaux pour défendre leurs opinions se soit appelé les **« Anonymous »** et adopte comme emblème un masque de carnaval.

Plus généralement, on peut revendiquer l'anonymat pour **protéger sa vie privée** et agir selon son bon vouloir, sans contrainte. L'anonymat apparaît alors comme le *garant de la liberté* entendue comme la possibilité de se comporter à sa guise. Cela porte entre autres sur les traces que nous laissons, à tout moment, lorsque nous payons avec nos cartes de crédit, voyageons avec nos passes numériques, utilisons nos téléphones portables, appelons nos correspondants, allons chez le médecin, achetons des médicaments, etc.

2.3 Limites de l'anonymat

Nous distinguons des limites propres à la nature des données et des limites d'ordre éthique.

2.3.1 Limites intrinsèques

La première limite tient à l'existence de données qui sont si intimement liées à la personne qu'il est impossible, avec elles, de conserver l'anonymat. C'est en particulier le cas de données biométriques en

2. LOI n° 2021-1017 du 2 août 2021 relative à la bioéthique. `https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000043884384`

3. Selon l'Art. L. 2143-2 de cette loi, « Toute personne conçue par assistance médicale à la procréation avec tiers donneur peut, si elle le souhaite, accéder à sa majorité à l'identité et aux données non identifiantes du tiers donneur ».

4. *L'Épineux dossier des machines à voter* Rapport N° 445 du Sénat (9 avril 2014) cf. `http://www.senat.fr/rap/r13-445/r13-4451.pdf`

général, par exemple des visages, des empreintes digitales, des postures comme le comportement sur un clavier ou la démarche, ou des données génomiques. Dans ce dernier cas, la séquence ADN identifie non seulement l'individu, sans équivoque, mais aussi l'ensemble de son lignage, à savoir son ascendance et sa descendance. Nous verrons, dans la suite du rapport, les précautions qu'il convient de prendre avec ce type de données.

2.3.2 Limites éthiques

La possibilité, relativement récente dans l'histoire, de mémoriser, d'accéder et de traiter de grandes quantités de données nominatives, non nominatives ou anonymisées renouvelle un dilemme ancien, celui de l'anonymat, et nous met face à un dilemme nouveau, celui de l'anonymisation.

2.3.2.1 Dilemme de l'anonymat

Le dilemme de l'anonymat peut s'illustrer par un exemple. Faut-il permettre aux utilisateurs des réseaux sociaux et des plateformes de blogage ou de micro-blogage, de diffuser des messages en restant anonymes ou en utilisant un pseudonyme, c'est-à-dire sans dévoiler leur identité ? De même, faut-il permettre à une personne d'envoyer un courriel ou de téléphoner sans dévoiler son identité ?

Répondre positivement à cette question, revient à permettre aux émetteurs de fausses informations de nous manipuler et aux trolls [5] d'injurier en toute impunité.

Mais y répondre négativement enjoint les utilisateurs à une transparence totale, injonction parfois justifiée par l'argument (simpliste) que ceux qui n'ont rien à cacher n'ont rien à craindre. Par exemple, un salarié militant dans une association ne pourrait pas utiliser un pseudonyme pour ses activités militantes, et ainsi s'assurer que son employeur n'en a pas connaissance. Il se trouverait alors dans l'impossibilité de cloisonner les contextes dans lesquels il entend évoluer : contexte professionnel, familial, personnel, associatif [6].

Notons que l'aspiration à l'anonymat se heurte non seulement à des difficultés intrinsèques qui tiennent à des obstacles techniques (par exemple, données qui trahissent directement l'identité de l'individu, comme les données génomiques), mais aussi à des considérations éthiques qui font que dans certains cas de figure, il est souhaitable, de dévoiler des informations sur les personnes, sans rien masquer.

Dans le contexte juridique, certains revendiquent une liberté « de ne pas apparaître », qui prendrait la forme d'un droit à l'anonymat dès lors que celui-ci est « un moyen d'éviter les représailles ou l'attention non voulue, [...] de nature à favoriser grandement la libre circulation des informations et des idées, notamment sur internet » [7]. Préservé à titre de principe, l'anonymat serait levé ponctuellement, notamment à des fins de poursuites judiciaires. [8]

Pour l'heure toutefois, la reconnaissance d'un tel droit demeure encore parcellaire : elle transparaît notamment de l'encadrement de l'activité des fournisseurs d'accès à internet, tenus d'anonymiser les données relatives au trafic ou de respecter de stricts délais quant à la conservation des données [9]. Cette

5. Un troll est un internaute qui écrit de manière intentionnelle des messages désobligeants, polémiques, provocants, absurdes, de mauvaise foi, voire insultants, et souvent répétitifs, sur des sites communautaires et de dialogues tels que les forums de discussion. Wikipedia le 22 octobre 2021.

6. Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2010.

7. CEDH, 16 juin 2015, Delfi AS c. Estonie, req. n° 64569/09, pt. 147, Comm. com. électr. 2015, comm. n° 68, obs. Loiseau G. ; RTDH 2016, n° 108, p. 954, obs. Montero et Van Enis. Voir : `https://hudoc.echr.coe.int/eng?i=001-155627`.

8. Sur la consécration embryonnaire d'un tel droit voir Pierre Leclerq. « L'anonymat : une situation souvent légitime ; rarement un droit ». In : *Droit et technique : études à la mémoire du professeur Xavier Linant de Bellefonds*. Litec, impr., 2007. isbn : 978-2-7110-0641-0.

9. CPCE (Code des postes et des communications électroniques), art. L. 34-1-I – *AddeCJUE, Digital Rights Ireland, préc.*

réflexion se prolonge s'agissant de la protection du chiffrement, notamment en raison des atteintes à la protection des données personnelles et au secret des correspondances qui peuvent résulter d'une surveillance massive [10].

2.3.2.2 Dilemme de l'anonymisation

Le dilemme de l'anonymisation est d'une nature différente.

Les masses de données accumulées sur les personnes par les hôpitaux, écoles, administrations... peuvent s'avérer d'une grande utilité, en particulier pour les chercheurs. Mais diffuser, par exemple des dossiers médicaux, même en direction de chercheurs, soulève des questions éthiques, en particulier parce que ces chercheurs, même s'ils ne publient pas directement ces données, publieront des informations qui en sont issues. Les professionnels de santé qui communiquent ces dossiers aux chercheurs doivent donc veiller à ce que ces informations ne permettent pas, à des individus mal intentionnés, de remonter aux données initiales et d'en déduire l'état de santé d'une personne particulière.

Une solution pour cela consiste à anonymiser les données, ce qui implique non seulement de supprimer le nom des patients, mais aussi de modifier les informations de manière à empêcher la ré-identification (voir chapitre 4).

La question du degré de modification de ces informations constitue le dilemme de l'anonymisation :

— Si on les transforme trop peu — dans le cas limite si on transmet le dossier entier — on rend la ré-identification possible,

— Si on les transforme trop — dans le cas limite, on vide complètement le dossier — on les rend moins pertinentes pour la recherche, voire inutiles.

Nous ne savons pas aujourd'hui garantir simultanément le parfait anonymat des patients et la meilleure efficacité de la recherche. Nous devons donc sacrifier une petite partie de l'un au profit de l'autre : des choix doivent être effectués, choix qui peuvent varier en fonction du contexte de la recherche.

Ce dilemme est nouveau, car, naguère, l'état des techniques ne permettait pas la collecte massive de données sur les patients, élèves ou citoyens et aujourd'hui on dispose de capacités de traitement et de ré-identification considérablement accrues.

2.3.2.3 Dilemme de la minimisation

La volonté de minimiser le recueil et le stockage des données personnelles entre en tension avec l'esprit même de l'exploitation de grandes masses de données (*Big Data*), qui tire avantage de leur quantité.

D'un côté, tant pour la protection de la vie privée des personnes, avec le RGPD, que pour des raisons de sobriété numérique, il apparaît souhaitable de limiter, dans la mesure du possible, le nombre et la taille des données conservées et traitées. Cela a conduit à une volonté de «minimisation des données» que traduisent deux principes, le «principe de finalité» selon lequel des données personnelles ne peuvent être recueillies sans que l'on ait auparavant défini et explicité clairement les objectifs poursuivis, et le «principe de proportionnalité» au nom duquel on doit spécifier la quantité de données nécessaire au regard de la finalité du recueil ainsi que la durée au delà de laquelle les données personnelles doivent être détruites.

D'un autre côté, plus le nombre et la taille des données s'accroissent, meilleurs pourraient être les résultats. La poursuite des objectifs technologiques conduit donc à tout mettre en œuvre pour augmenter, autant que faire se peut, la quantité des données stockées et traitées. À cette fin, on souhaite tirer parti de leur ouverture et de leur conservation, pour les accumuler, sans savoir *a priori* l'utilisation que l'on en fera.

10. CERNA. *La souveraineté à l'ère du numérique : Rester maîtres de nos choix et de nos valeurs*. fr. Rapp. tech. Nov. 2018, p. 36. URL : http://cerna-ethics-allistene.org/digitalAssets/55/55708_AvisSouverainete-CERNA-2018.pdf.

Il en résulte une tension entre deux objectifs antagoniques, satisfaire le principe de minimisation des données et accroître la qualité des résultats par des moyens technologiques. La difficulté tient à ce qu'on n'est pas en mesure, au moment de la collecte, de savoir comment les données seront utilisées *in fine*.

2.3.3 Données ouvertes et anonymisation

On pourrait penser qu'en pratique, l'anonymat des données n'est pas nécessaire dans le contexte de la recherche, à condition que les chercheurs ayant connaissance de données à caractère personnel, s'engagent à n'en faire usage que pour les besoins de leurs recherches ET que celles-ci ne visent que l'avancée de connaissances générales. Il n'y aurait alors rien à craindre pour la vie privée de l'individu. Autrement dit, les données de recherche auraient au sens juridique un statut « exorbitant », en cela qu'il sortirait du droit commun. Cela signifie que les chercheurs entreraient dans une forme de confidentialité, impliquant qu'ils seraient responsables de la mauvaise utilisation qu'ils pourraient faire des données personnelles, mais aussi qu'ils ne les communiqueraient pas.

La science ouverte et la volonté de partager les données de recherche ajoutent un degré de complexité. Dans le cadre de la science ouverte, il est aujourd'hui demandé que les travaux issus de recherches financées sur des fonds publics national ou européen soient en accès libre. Cela va de pair avec la volonté de partage des données de recherche. Outre son caractère généreux et la capacité scientifique fondamentale de pouvoir vérifier une assertion, ce partage peut se justifier au plan technique par la capacité de traiter des masses de données importantes avec les techniques d'apprentissage machine, et surtout par le fait que la qualité des résultats dépend souvent de la quantité d'exemples. Il s'ensuit qu'il devient utile de partager les données de recherche, y compris les données à caractère personnel. Or, avec ce partage des données à caractère personnel et cette ouverture des données, il n'y a plus de confidentialité possible. Une solution envisageable consiste donc à essayer de rendre les données anonymes, d'où l'importance de techniques qui permettent de le faire de façon fiable, sans possibilité de remonter aux personnes qui sont à l'origine des données et sans trop dégrader ces mêmes données.

2.3.4 Consentement et anonymat

Le consentement libre et éclairé a été progressivement introduit dans la bioéthique pour les expérimentations médicales d'abord [11], puis plus tard pour tous les soins [12]. Il part du principe selon lequel le « patient » décide lui-même, après avoir été informé, des traitements et des interventions qu'il subit. Cela traduit le principe d'autonomie de la personne humaine : elle choisit par elle-même ce qui la concerne. Dans le contexte de la bio-éthique [13], il existe une tension entre le principe de bienfaisance, qui voudrait que les médecins imposent la conduite qu'ils pensent optimale, et le principe d'autonomie de la personne, qui lui laisse le libre choix. On doit ajouter que, dans le cas des soins, certains y voient une façon de dé-responsabiliser le médecin qui se décharge des conséquences de ses décisions sur le patient. Enfin, dernier point essentiel ici, le consentement doit être *révocable à tout moment*.

Venons en maintenant au cas des données personnelles et du renoncement à l'anonymat. La législation sur l'utilisation des données personnelles se calque sur celle de la bioéthique : on doit donner son consentement libre et éclairé, en cela qu'on doit savoir à quoi servent les données et qui les utilise. De plus, il est loisible de retirer à tout moment son consentement.

11. Voir rapport de Belmont `https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html`
12. voir la LOI n° 2002-303 du 4 mars 2002 relative aux droits des malades et à la qualité du système de santé, `https://www.legifrance.gouv.fr/eli/loi/2002/3/4/2002-303/jo/texte`
13. Voir l'avis n°58 du CCNE intitulé *Consentement éclairé et information des personnes qui se prêtent à des actes de soin ou de recherche (1998-06-12)* et paru le 12 juin 1998, `https://www.ccne-ethique.fr/sites/default/files/2021-02/avis058.pdf`

Toutefois, dans le cas des données, le retrait apparaît plus difficile, car la duplication des données et leur divulgation peuvent se faire sans que la personne n'ait été informée. Dans cette dernière éventualité, il n'est plus possible de savoir qui utilise les données ; le consentement et *a fortiori* le retrait de consentement se trouvent mis en défaut. À cela, on doit ajouter que le consentement aux CGU (Conditions générales d'utilisation) n'est pas tout à fait libre, car il est rare qu'on ait le choix, ni tout à fait éclairé, parce qu'elles sont cachées dans des textes très longs et peu explicites.

Il convient maintenant de se demander ce que pourrait signifier renoncer à son anonymat pour savoir si, dans ce cas, un consentement serait possible. En effet, puisque toute donnée personnelle permet, éventuellement, de remonter à la personne, cela pourrait vouloir dire que l'on accepte que ses données personnelles soient utilisées, ce qui recouvrirait des situations décrites dans le RGPD. Dans cette éventualité, il n'y aurait rien de plus à ajouter que ce que dit le RGPD. Mais, cela peut signifier que l'on accepte de diffuser des données personnelles identifiantes. Dans cette dernière éventualité, on ne disposerait plus d'aucun droit de retrait, car rien n'empêcherait que les données aient été dupliquées et divulguées. Il s'agirait donc d'un consentement atypique dont il faudrait que les personnes prennent conscience.

On ne peut donc pas consentir, au sens de la bioéthique, à renoncer à son anonymat, en raison de l'impossibilité de révocation d'un tel consentement. Il ne serait donc pas éthique de le demander. De façon similaire, chercher à ré-identifier pour son propre usage scientifique des données personnelles ne serait pas intègre, à moins que ceci ne soit validé par un comité opérationnel d'éthique compétent.

Recommandation pour les Scientifiques - 1

— Il ne serait pas éthique de demander, dans un protocole de recherche, à un sujet humain de consentir à renoncer à son anonymat. En général, il ne serait pas intègre de chercher à ré-identifier pour son propre usage scientifique des données personnelles.

2.4 Définitions et mise en contexte

Après ces préliminaires, et avant d'entrer dans le vif du sujet, précisons, au risque de répéter un peu qui a déjà été dit, la signification de quelques termes gravitant dans la nébuleuse terminologique du concept d'anonymat, en particulier l'adjectif anonyme et les notions d'anonymisation des données, d'identité, d'identités multiples, d'usurpation d'identité, d'avatar, de pseudonyme, de confidentialité, d'authentification et d'identification.

Anonyme : étymologiquement « sans nom », renvoie à un objet dont on ne connaît pas l'auteur. Ainsi en va-t-il d'une lettre anonyme, d'un tableau ou, plus généralement, d'une œuvre anonyme, ou encore, et c'est la chose qui nous intéresse ici, de données anonymes. Notons que de tout temps cette définition du terme « anonyme » sous-entend que l'*on* n'est pas capable de déterminer l'*auteur*, à un *moment* donné et avec des *moyens* légaux et raisonnables. Le numérique modifie fortement ces éléments (qui, quand, comment) du fait de l'accroissement des capacités de calcul.

Anonymat : qualité de ce qui est anonyme.

Données personnelles : on qualifie de personnelles toutes les données en lien avec une personne (lien de production, de statut, d'appartenance, de référence etc.).

Données sensibles : les données sensibles sont des données que l'on ne peut ni recueillir, ni traiter sans autorisation, ni communiquer largement. Parmi elles, on distingue les données à caractère personnel des données stratégiques pour l'entreprise [14] ou pour l'État. Nous ne traiterons ici que des *données personnelles sensibles*.

Les données personnelles dites sensibles sont en France des données liées à l'origine raciale ou ethnique, à des opinions politiques, des convictions religieuses ou philosophiques ou à l'appartenance syndicale, ou des données relatives à la santé et à la vie sexuelle, ou encore les données relatives aux infractions

14. https://www.journaldunet.com/management/direction-generale/1422246-secret-des-affaires-une-loi...

pénales et aux condamnations et plus récemment, avec l'article 9 du RGPD, des données biométriques ou des données génétiques.

Données anonymes et ***données de source anonyme*** : lorsque l'adjectif anonyme qualifie des données, cela possède deux sens distincts. En se référant à l'étymologie, et par analogie à une lettre ou une œuvre anonyme, cela désigne des données dont *on ne connaît pas l'auteur* ; nous parlerons alors de ***données de source anonyme*** ; quand il s'agit de données à caractère personnel, autrement dit de données qui portent sur des personnes, par exemple de données médicales, cela qualifie des *données qui ne permettent pas de remonter aux personnes qu'elles décrivent*. Dans cette dernière éventualité, nous parlerons simplement de ***données anonymes***.

En écho à ces deux définitions de l'anonymat des données que nous avons mentionnées, l'**anonymisation** renvoie à deux démarches différentes : indexanonymisation|textbf

— L'***anonymisation de la source des données,*** c'est-à-dire *le processus qui consiste à rompre le lien entre les données et ceux qui les ont produites*, ce qui permet par exemple d'empêcher de savoir qui a généré des infox, et

— L'***anonymisation des données***, c'est-à-dire le processus qui consiste à *modifier des données à caractère personnel pour, à un moment et dans un contexte donné, ne plus pouvoir faire le lien entre ces données modifiées et les personnes qu'elles décrivent*.

Dans le premier cas, cela recouvre toutes les techniques qui empêchent de remonter à la source d'une diffusion d'information, qu'il s'agisse d'une lettre, d'une *infox* sur la toile ou d'une requête sur un moteur de recherche. Dans les romans policiers du XX^{e} siècle, on découpait des caractères typographiques dans un journal, puis on les collait, ou on utilisait une machine à écrire en libre accès. Aujourd'hui, sur internet, on utilise le serveur TOR, même si cela ne prémunit pas totalement contre une enquête approfondie qui tracerait l'origine des paquets. Enfin, on utilise des moteurs de recherche comme Qwant ou DuckDuckGo pour éviter que l'analyse de nos requêtes ne permette de nous identifier, ce qui correspond aussi à une forme d'anonymat sur le web.

Dans le second cas, pour éviter de retrouver la personne décrite par des données, on peut être tenté, pour procéder à une anonymisation, de se contenter d'éliminer les références immédiates à la personne, en particulier son nom. Or, même lorsqu'il est possible d'éliminer le nom, on peut le recouvrer par croisement avec d'autres bases de données. Ainsi, en 1997, la base de données médicales GIC (Group Insurance Commission) aux USA dans laquelle on avait simplement supprimé le nom, le prénom et le numéro de sécurité sociale des individus, mais gardé le code postal, le genre et la date de naissance, ne pouvait pas être considérée comme anonymisée, puisque après croisement avec les listes électorales qui comprenaient également ces trois informations mais aussi le nom, un chercheur avait pu identifier le dossier médical du gouverneur du Massachusetts [15]. Pour plus de détails sur la ré-identification, voir section 4.1.

Confidentialité : la *confidentialité* repose sur la *confiance dans une personne ou une institution (organisme public, journal, etc.) qui relaie ou traite des informations personnelles dont elle se porte garante, sans en dévoiler l'origine*. À titre d'illustration de tels régimes de confidentialité, citons l'accouchement sous X, la protection du « secret des sources » que revendiquent les journalistes, etc.

Nous avons défini l'anonymat comme l'absence de lien entre des données et une personne, qu'il s'agisse de la personne qui a produit ces données ou de la personne que décrivent les données. Mais, comment repérer une personne, en dépit de ces transformations variées ? C'est l'objet des procédures d'identification. Plus précisément, on définit les notions suivantes :

Identité : au sens général, l'identité désigne à la fois ce qui rapproche des entités distinctes (individu, objet ou collectivité) ou ce qui appartient à une seule entité (ainsi peut-on parler d'une identité régionale

15. L. Sweeney. « Weaving technology and policy together to maintain confidentiality. » eng. In : *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics* 25.2-3 (1997). Place: England, p. 98-110, 82. issn : 1073-1105. doi : 10.1111/j.1748-720x.1997.tb01885.x.

ou nationale) ou encore ce qui dans une entité demeure dans le temps. Nous nous restreindrons ici uniquement aux personnes et à ce qui perdure, à savoir, au *caractère de ce qui, sous les évolutions variées d'âge, d'habits, de domicile etc., demeure le même*. Au sens juridique, l'identité des individus recouvre un « ensemble des traits ou caractéristiques qui, au regard de l'état civil, permettent de reconnaître une personne et d'établir son individualité au regard de la loi » [16]. *A priori*, le nom et le prénom devraient suffire. Cependant, compte tenu des risques d'homonymie, on ajoute souvent la date et le lieu de naissance. Les États disposent d'autres éléments d'identification, par exemple, en France le Numéro d'Inscription au Répertoire (NIR) [17] ou des empreintes digitales. On peut ajouter d'autres éléments d'ordre biologique, ou des éléments de posture, comme le comportement sur un clavier ou la démarche, ou encore le téléphone portable des individus.

Identification (ang : *identification*) : par *identification*, en entend un processus de découverte de l'identité de la personne.

Authentification (ang : *authenticate)* : par *authentification*, on entend un processus de vérification de l'identité d'une personne à partir de données complémentaires.

Prendre la photographie du visage d'un piéton en pleine rue et retrouver son identité correspond à une identification ; en revanche, s'assurer, avec le système PARAFE [18], que le possesseur du passeport correspond bien à la personne qui passe dans le sas relève de l'authentification.

Identifiant : un *identifiant* est une propriété, une donnée ou un objet qui permet d'identifier une personne.

Identifier : procéder à l'identification.

Quasi-identifiant : dans une base de données personnelles relationnelles, un *quasi-identifiant* est un ensemble d'attributs d'une table relative à des personnes pour lesquels il existe au moins une combinaison de valeurs qui n'apparaît qu'une seule fois dans cette table. Cette combinaison permet alors d'identifier ces personnes ou au moins l'une d'entre elles [19].

Profil : étymologiquement, le profil désigne la vue de côté d'un visage. On en a dérivé une vue simplifiée des choses. Le terme est utilisé en architecture, dans l'industrie (le profil mécanique d'une pièce), en géographie (le profil d'un cours d'eau), en médecine (le profil d'un malade), en psychologie et en criminologie. Nous nous intéressons ici à ces trois dernières acceptions, car elles se rapportent toute à un ensemble de traits d'une personne.

Profilage : processus qui permet d'établir un profil. Comme le terme profil désigne une vue simplifiée et qu'il s'applique donc à de multiples choses, cela recouvre des procédés très variés. Nous nous intéresserons ici au profilage de la personne.

Ce terme vient de l'analyse comportementale. Il a été introduit dans les méthodes d'investigation policières par le docteur Thomas Bond au XIX^{e} siècle pour caractériser la psychologie des suspects et établir leurs points faibles. Il a été repris par les spécialistes du marketing puis par des informaticiens pour automatiser le processus à partir d'informations personnelles [20]. Il joue maintenant un rôle central dans l'économie d'internet pour cibler notamment les publicités.

Pseudonyme et ***identités multiples*** : si l'on se limitait à la caractérisation de l'identité par le nom, il

16. Rédigé par le Réseau des Référents Régionaux d'IdentitoVigilance (3RIV), le référentiel national d'identitovigilance est composé de plusieurs parties différenciées par thèmes. La première partie (RNIV-1) présente les principes communs d'identification des usagers.

17. Le NIR est un identifiant de 15 chiffres décimaux, aussi appelé numéro Insee ou encore numéro de sécurité sociale. Le répertoire national d'identification des personnes physiques (RNIPP) est un répertoire français tenu par l'Insee, recensant les personnes vivantes ou décédées.

18. https://www.immigration.interieur.gouv.fr/Europe-et-International/La-circulation-transfrontiere/Le-passage-rapide-aux-frontieres-exterieures-PARAFE

19. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, http://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf

20. https://gdpr-info.eu/art-4-gdpr

suffirait d'adopter un *pseudonyme*, c'est-à-dire un *nom d'emprunt*, pour échapper à une identification et demeurer anonyme, ou plus exactement, se démultiplier en différentes personnes indépendantes. C'est d'ailleurs ce que font beaucoup de personnes sur les réseaux sociaux et, plus généralement, sur internet, pour éviter de se dévoiler. Et, comme il est possible de multiplier les pseudonymes, chacun étant associé à une facette différente de sa personne qui peut dès lors se présenter sous des jours variés à des publics variés (amis, famille, milieu professionnel, etc.) on parle parfois d'*identités multiples*, ce qui relève quelque peu de l'oxymore puisque cela signifie que le même apparaît différent...

Pseudonymisation : processus de traitement des données à caractère personnel qui consiste à remplacer les noms par des pseudonymes. Par extension, en particulier dans le RGPD (voir section 3.1.2), on appelle pseudonymisation le processus de remplacement de tous les identifiants réalisé de manière à ce qu'on ne puisse plus attribuer les données à une personne physique sans information supplémentaire intentionnellement conservée [21]. Dans ce dernier sens, la pseudonymisation est synonyme de la désidentification (voir *infra*). En cas de besoin, la pseudonymisation permet de retrouver l'identité des personnes avec les informations supplémentaires conservées à cette fin.

Désidentification (ang : *de-identification*) : processus de traitement des données à caractère personnel qui consiste à remplacer les identifiants de façon à ce que l'on ne soit plus en mesure de les rapporter à une personne sans information supplémentaire. Ce terme est surtout utilisé dans le traitement de données textuelles (cf. section 4.2.2.1).

Des auteurs [22] déplorent un certain flou dans la terminologie utilisée par certains chercheurs dans le domaine biomédical, venant d'une confusion avec l'anonymisation d'un côté et la pseudonymisation de l'autre.

Sur ce dernier point, tandis que le RGPD use du terme pseudonymisation, il apparaît que le terme désidentification est employé dans la législation américaine où il est défini dans la HIPAA (*American Health Insurance Portability and Accountability Act*, 1996b [23]) §164.514(a) et (b). Aux États-Unis, on appelle PHI (*Personal Health Identifier*) un identifiant de santé personnel qu'on peut trouver dans des documents cliniques. Le HIPAA a répertorié 18 PHI qui doivent être retirés pour que le texte soit considéré comme désidentifié.

Quant à la distinction avec l'anonymisation, elle tient à ce que la désidentification comme la pseudonymisation signifie que l'on remplace les identifiants, tandis que l'anonymisation indique qu'en modifiant les données personnelles on n'est plus en mesure de remonter à la source.

Avatar : Au-delà des différents noms, il est possible d'adopter différents attributs, associés à chacune des identités multiples que l'on se construit. Cela correspond à la notion d'*avatar* qui désignait, initialement, chacune des différentes incarnations du dieu hindou Vishnou, et qui recouvre maintenant chacune des personnalités qu'un individu choisit d'adopter dans le monde numérique, avec un nom, une image, un sexe, une histoire, etc. qu'il se construit.

Usurpation d'identité : En contrepartie de la facilité avec laquelle un individu se démultiplie dans l'univers de la toile, apparaissent de nouvelles vulnérabilités, en particulier l'usurpation d'identité, lorsque des individus mal intentionnés s'emparent délibérément des caractéristiques d'une personne, sans son consentement, et en abusent pour la spolier, par exemple pour faire des achats à sa place. L'usurpation d'identité est une infraction sanctionnée par le code pénal : article 226-1-4 du Code pénal qui punit,

21. Le RGPD définit la pseudonymisation dans l'article 4, item 5, comme étant *le traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que ces données à caractère personnel ne sont pas attribuées à une personne physique identifiée ou identifiable.*

22. Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C, "Use and Understanding of Anonymization and De-Identification in the Biomedical Literature : Scoping Review", *J Med Internet Res*, 2019 ;21(5) :e13484, DOI : 10.2196/13484

23. `https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html`

depuis la loi du 14 avril 2011 n°2011/447 dite « LOPPSI 2 », « le fait d'usurper l'identité d'un tiers ou de faire usage d'une ou plusieurs données de toute nature à permettre de l'identifier en vue de troubler sa tranquillité ou celle d'autrui, ou de porter atteinte à son honneur ou à sa réputation » en le sanctionnant d'une peine d'un an d'emprisonnement et de 15000 euros d'amende.

2.5 Approche juridique

2.5.1 Information anonyme et donnée personnelle

Les informations **anonymes** sont considérées par le RGPD comme : « **les informations ne concernant pas une personne physique identifiée ou identifiable** » [24].

A contrario, toute information concernant une personne physique identifiée ou identifiable n'est pas une information anonyme ; cette information est alors qualifiée de donnée à caractère personnel. Ainsi le RGPD considère « toute information se rapportant à une personne physique identifiée ou identifiable » comme **une donnée à caractère personnel** [25].

2.5.2 La notion de personne « identifiable »

Si la notion de personne « identifiée » ne pose pas de problème particulier, celle de personne « identifiable » en revanche est plus délicate. Sur ce point primordial, qui permet de délimiter la frontière entre information anonyme et donnée à caractère personnel, le texte précise qu'une personne identifiable est une personne « qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'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 » [26]. Ainsi, ces types de données ne rentrent pas dans la catégorie des informations anonymes.

Il en va de même lorsque « l'ensemble des moyens raisonnablement susceptibles d'être utilisés par le responsable du traitement ou par toute autre personne » permettent d'identifier une personne physique, de manière directe ou indirecte [27]. Parmi ces moyens, le RGPD indique le ciblage (ang : *singling out*). Mais que doit-on entendre par « moyens raisonnablement susceptibles d'être utilisés » ? Sur ce point, l'analyse, car il s'agit ici d'une véritable analyse devant être menée au cas par cas, doit considérer « l'ensemble des facteurs objectifs, tels que le coût de l'identification et le temps nécessaire à celle-ci, en tenant compte des technologies disponibles au moment du traitement et de l'évolution de celles-ci » [28]. Par exemple, les clauses de protection des données personnelles du compteur Linky ont été jugées satisfaisantes par la CNIL [29] [30] [31].

24. Considérant 26 du Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE : JOUE, L 119 du 4 mai 2016, p. 1 (RGPD).
25. Article 4-1 du RGPD.
26. Article 4-1 du RGPD.
27. Considérant 26 du RGPD, `https://gdpr-text.com/fr/read/recital-26/`.
28. *Ibidem.*
29. `https://www.lemonde.fr/economie/article/2021/05/06/compteurs-linky-la-cnil-cloture-la-procedure-de-mise-en-demeure-d-engie_6079358_3234.html`
30. `https://www.cnil.fr/fr/linky-gazpar-quelles-donnees-sont-collectees-et-transmises-par-les-compteurs-communicants`
31. `https://www.lemonde.fr/les-decodeurs/article/2018/10/23/linky-en-questions-le-compteur-electrique-est-il-un-espion_5373380_4355770.html`

2.5.3 Caractère absolu/relatif de l'anonymisation

Il découle de ce qui vient d'être dit qu'il existe une zone de flou entre les données personnelles et les informations non-anonymes, qui tient à la difficulté à caractériser précisément la notion de « personne identifiable ». En effet, celle-ci dépend de la puissance des moyens technologiques mis en œuvre et des informations complémentaires dont on dispose.

2.5.3.1 Aspects contextuel et temporel

Ainsi, selon le contexte, selon les technologies utilisées, selon l'ensemble des informations figurant dans une base de données, une information peut être considérée comme anonyme à un instant *t*, et perdre ce caractère ultérieurement pour devenir une donnée à caractère personnel au sens du RGPD. Les progrès de l'informatique quantique, par exemple, permettent d'envisager que dans quelques années, on puisse déchiffrer en très peu de temps les informations chiffrées avec certains des moyens actuels. Il est donc raisonnable de prévoir que des méthodes informatiques de ré-identification plus puissantes seront développées. Il faut également prendre en compte le fait que le volume des données disponibles que l'on peut croiser entre elles s'accroît très rapidement. Ainsi, lorsque l'on considère une base de données décrivant des personnes non (encore) identifiées, on ne peut jamais être certain qu'il n'existe pas et *a fortiori* qu'il n'existera pas une autre base de données complémentaires qui permette ou permettra de ré-identifier les personnes.

Recommandation pour les Scientifiques - 2

— L'anonymisation n'est pas absolue et son efficacité dépend de facteurs évoluant dans le temps, comme la puissance des machines et les données disponibles. Il est nécessaire pour les scientifiques, les responsables de projets et d'unités de recherche d'adopter une approche dynamique, notamment en mettant en place une organisation interne appropriée afin de revoir régulièrement les techniques d'anonymisation et/ou de pseudonymisation utilisées au regard des avancées technologiques, de l'évolution des finalités poursuivies et des éventuelles catégories de données ajoutées dans un travail de recherche.

Recommandation pour les institutions de recherche et d'enseignement - 1

— Etant donné le caractère relatif et évolutif de l'anonymisation, il est souhaitable de se donner la capacité à renforcer la qualité et la sécurité des moyens de gestion des données mis à disposition des chercheurs par des mesures telles que :
(i) Renforcer et soutenir des initiatives qui visent à promouvoir les logiciels libres en développant des clouds maîtrisés ;
(ii) Favoriser des solutions basées sur des systèmes interopérables pour une recherche fédérée ;
(iii) Contribuer à mettre en place des infrastructures de collaboration et de mutualisation des données personnelles à des fins de recherche aux niveaux national, européen et international, selon les domaines de recherche, qui intègrent la protection de la vie privée.

2.5.3.2 Qui décide ? Sur quoi peut-on se baser ?

La certification de la conformité des processus d'anonymisation des données personnelles dans la perspective de leur mise en ligne et de leur réutilisation est devenue une mission de la CNIL [32], en vertu de la loi pour une République numérique du 7 octobre 2016 [33].

32. https://www.cnil.fr/fr/ce-que-change-la-loi-pour-une-republique-numerique-pour-la-protection-des-donnees-personnelles.

33. https://www.legifrance.gouv.fr/jorf/id/JORFARTI000033203249 et https://www.legifrance.gouv.fr/loda/article_lc/LEGIARTI000037090040/2018-05-25

Si certaines entreprises qualifient les procédés qu'elles utilisent de techniques d'anonymisation, leur approche jusqu'à présent n'a jamais été certifiée par la CNIL. En témoigne, par exemple, l'affaire JC-Decaux portant sur un traitement de données ayant pour finalité de tester une méthodologie d'estimation quantitative des flux piétons. Dans cette affaire, la CNIL a considéré que le procédé utilisé ne constituait pas une technique d'anonymisation mais une technique de pseudonymisation [34], analyse qui a été validée par le Conseil d'Etat [35].

34. Délibération de la Cnil n°2015-255 du 16 juillet 2015 refusant la mise en œuvre par la société JCDecaux d'un traitement automatisé de données à caractère personnel ayant pour finalité de tester une méthodologie d'estimation quantitative des flux piétons sur la dalle de La Défense (demande d'autorisation n° 1833589).

35. Conseil d'Etat, 10ème - 9ème chambres réunies, 08/02/2017, 393714, https://www.legifrance.gouv.fr/affichJuriAdmin.do?idTexte=CETATEXT000034017907.

3 Signification et principes d'anonymisation pour la recherche scientifique

3.1 Anonymiser et pseudonymiser au sens du RGPD

3.1.1 Anonymisation au sens du RGPD

On a vu en section 2.4 que le processus d'anonymisation des données à caractère personnel est le processus qui vise à rendre les données anonymes [1]. C'est un processus irréversible en ce sens que les résultats du traitement ne permettent pas de reconstituer l'intégralité des données d'origine.

Le RGPD dans son considérant 26 précise en quoi l'obligation de résultat, obtenir des données anonymes, est en fait une obligation de moyens. Rappelons les points que le RGPD demande de prendre en compte pour savoir si des données sont correctement anonymisées, c'est-à-dire, non susceptibles d'être ré-identifiées avec des moyens raisonnables [2] : la présence d'informations supplémentaires disponibles ; le coût et le temps nécessaire pour le processus de ré-identification ; les technologies disponibles au moment du processus.

Recommandation pour les institutions de recherche et d'enseignement - 2

— Expliciter pour les scientifiques la portée du Considérant 26 du RGPD relatif aux données anonymes et à la notion de personne physique identifiable.

Un processus d'anonymisation constitue ou fait partie d'un traitement de données à caractère personnel dans le sens où des données à caractère personnel sont utilisées en entrée de la fonction d'anonymisation. En revanche, le RGPD spécifie qu'il ne s'applique pas aux données issues d'un processus d'anonymisation. En effet, d'après le considérant 26 du RGPD « Il n'y a pas lieu d'appliquer les principes relatifs à la protection des données aux informations anonymes, à savoir les informations ne concernant pas une personne physique identifiée ou identifiable, ni aux données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable. »

Cette anonymisation peut intervenir soit dans un bref délai, c'est-à-dire suivre immédiatement la collecte des données, soit ultérieurement. La loi Informatique et Libertés prévoyait une telle distinction bref / long délai de traitement. Il en résultait des facilités de mise en œuvre quand on se trouvait dans le premier cas (avec par exemple la possibilité d'être exempté du droit d'opposition). Ce n'est plus le cas avec le RGPD puisque l'anonymisation est un traitement de données à caractère personnel. Néanmoins, on peut souligner l'intérêt d'un point de vue « respect de la vie privée » de réaliser une anonymisation à bref délai lorsque cela est possible puisque c'est une mesure protectrice.

3.1.2 Pseudonymisation au sens du RGPD

La pseudonymisation est un « traitement de données à caractère personnel de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires » [3]. A cet égard, ce texte exige que les informations établissant la correspondance entre

1. "... Données à caractère personnel rendues anonymes de telle manière que la personne concernée ne soit pas ou plus identifiable" (Considérant 26 du RGPD, `https://gdpr-text.com/fr/read/recital-26/`).
2. Comme développé dans les sections 2.5.2 et 2.5.3.
3. Article 4-5 du RGPD.

les noms et les pseudonymes soient conservées séparément et soumises à des mesures techniques et organisationnelles de protection.

Dans tous les cas, **les données pseudonymisées doivent être considérées comme des données à caractère personnel** concernant une personne identifiable, et ce, même si la table des correspondances avec les identifiants directs a été supprimée.

L'objectif recherché ici est de limiter le risque de ré-identification directe, par exemple en remplaçant un identifiant par un unique pseudonyme, ce qui permet alors l'étude de corrélations en cas de besoin particulier. La CNIL donne des exemples de pseudonymisation dans sa fiche sur les enjeux et avantages de l'anonymisation et de la pseudonymisation dans le cadre de la recherche scientifique (hors santé) [4] (voir aussi ci-dessous le chapitre 6 : *En pratique, en tant que scientifique, comment dois-je procéder ?*).

3.1.3 Intérêt légal de l'anonymisation et de la pseudonymisation

L'avantage du recours aux techniques d'anonymisation et de pseudonymisation est qu'elles participent à la réduction des risques à la fois pour les individus et pour les organismes collectant et traitant les données. Du côté des individus, elles contribuent à mieux protéger leurs droits fondamentaux, notamment le droit au respect de la vie privée et à la protection des données à caractère personnel. Du côté des organismes, elles aident les responsables du traitement et les sous-traitants à remplir leurs obligations, notamment les obligations de minimisation des données (seules des données adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées doivent être utilisées), de protection des données personnelles dès la conception (Data Protection by Design) et de sécurité des traitements.

3.2 RGPD, recherche scientifique et historique, statistiques

Le RGPD prévoit un cadre particulier pour les traitements poursuivant une finalité de recherche scientifique. Le texte définit la recherche scientifique au sens large qui comprend, par exemple, « le développement et la démonstration de technologies, la recherche fondamentale, la recherche appliquée et la recherche financée par le secteur privé » ainsi que « les études menées dans l'intérêt public dans le domaine de la santé publique » [5].

Il ne distingue pas la recherche publique et la recherche privée, ce qui constitue une différence par rapport aux traitements réalisés à des fins archivistiques. Pour ces derniers, le RGPD prend le soin de préciser qu'ils doivent être effectués « dans l'intérêt public ».

L'anonymisation ou la pseudonymisation doivent être mises en œuvre toutes les fois où elles s'avèrent pertinentes [6]. Afin de concilier l'impératif de protection des personnes physiques avec les spécificités de la recherche, des conditions particulières s'appliquent.

D'une part, les chercheurs bénéficient d'une dérogation au principe de limitation des finalités [7]. Ils peuvent se rapprocher des responsables de traitement ayant collecté de manière licite des données personnelles afin que celles-ci leur soient remises pour conduire leurs recherches. Ils ne sont donc pas obligés de collecter par eux-mêmes les données, notamment sur la base du consentement des personnes.

Ils peuvent aussi passer par des tiers afin de se faire confier des données par le responsable du traitement initial, puisque le changement de finalité à des fins de recherche est admis par le RGPD. On peut imaginer

4. Fiche CNIL Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation, janvier 2022. `https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation`

5. Considérant 159 du RGPD.

6. Article 89, §1 du RGPD.

7. Voir le considérant 50 et art. 5 du RGPD.

que ce type de partenariats de recherche passe par l'établissement de conventions aux termes desquelles le fournisseur de données et l'équipe de recherche se reconnaissent comme co-responsables du traitement.

D'autre part, ils disposent d'une certaine marge de manœuvre pour formuler les finalités des traitements de données collectées d'une manière moins précise que ce qui est exigé en principe par le texte [8]. Il est admis que cette finalité puisse s'élargir ou se préciser au fil du projet de recherche et en fonction des nécessités.

Par ailleurs, les données peuvent être conservées au-delà de la durée qui a été nécessaire pour atteindre la finalité de recherche (par exemple, au-delà de la durée d'un projet de recherche déterminé) du moment qu'elles sont conservées uniquement pour être utilisées à des fins de recherche [9], situation que l'on pourrait trouver par exemple dans l'exploitation de résultats de recherche sur des décisions de justice.

Notons que dans tous les cas, l'anonymisation est imposée lors de la diffusion des résultats de recherche, « *sauf* si l'intérêt des tiers à cette diffusion prévaut sur les intérêts ou les libertés et droits fondamentaux de la personne concernée » [10].

Recommandation pour les Scientifiques - 3

— Lorsque dans le cadre de ses travaux de recherche, un scientifique pense pouvoir s'appuyer sur l'Article 116 du décret n○ 2019-536 du 29 mai 2019, il devra s'entourer de l'avis écrit du comité opérationnel d'éthique et de celui du Délégué à la Protection des Données (DPD) (en anglais le DPO (*Data Protection Officer*)) de son établissement et s'assurer qu'il a obtenu toutes les autorisations requises par l'article 116 suscité [11].

8. Voir le considérant 33 du RGPD.

9. RGPD : « *Les données à caractère personnel peuvent être conservées pour des durées plus longues dans la mesure où elles seront traitées exclusivement à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques conformément à l'article 89, paragraphe 1, pour autant que soient mises en œuvre les mesures techniques et organisationnelles appropriées requises par le présent règlement afin de garantir les droits et libertés de la personne concernée (limitation de la conservation).* »

10. Article 116 du décret n° 2019-536 du 29 mai 2019 pris pour l'application de la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés : `https://www.legifrance.gouv.fr/loda/id/JORFTEXT000038528420`.

11. La diffusion de données à caractère personnel figurant dans des documents consultés en application de l'article L. 213-3 du code du patrimoine ne peut intervenir qu'après autorisation de l'administration des archives, après accord de l'autorité dont émanent les documents et avis du comité du secret statistique institué par l'article 6 bis de la loi n° 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques en ce qui concerne les données couvertes par le secret en matière de statistiques.

4 Les techniques d'anonymisation et leurs limites

4.1 Pseudonymisation et ré-identification

On a vu que le RGPD dans son considérant 26 (cf section 3.1) considérait des données pseudonymisées susceptibles d'être ré-identifiées avec des moyens raisonnables comme non anonymisées. C'est le cas si on peut croiser des données personnelles pseudonymisées avec un autre ensemble de données, permettant ainsi la ré-identification.

On a cité dans la section 2.4 l'exemple de la ré-identification du dossier médical du gouverneur du Massachusetts. Depuis, de nombreux travaux ont montré comment on peut ré-identifier en croisant plusieurs sources de données.

Nous allons illustrer par un exemple la façon dont des données uniquement pseudonymisées peuvent conduire, par croisement avec des données complémentaires, à une ré-identification, ce qui montre que ce sont toujours des données personnelles, comme il est dit dans la section 3.1.2.

Considérons le tableau de données A de la Figure 4.1 qui comporte des informations sensibles sur les pathologies dont souffrent un certain nombre de personnes. On a pseudonymisé les données en retirant les noms et prénoms et en rajoutant un pseudo arbitraire (première colonne). Mais on ne peut pas considérer que les données sont anonymisées car on peut les ré-identifier en les ***croisant*** avec une autre base publique B, parlant de ces personnes (ainsi qu'éventuellement d'autres personnes), et où les noms et prénoms, données directement identifiantes dans ce cadre d'étude, figurent explicitement.

On peut en effet facilement retrouver les noms et prénoms de certaines personnes du tableau A à partir du tableau B de la Figure 4.2, car certains individus sont identifiés dans B de façon unique par leur code postal d'adresse, leur genre et leur année de naissance, informations qu'on retrouve dans le tableau A. On dit que {CP_adr, Année_naiss, Genre} est un *quasi-identifiant* (défini formellement page 14) pour le tableau A, car c'est un ensemble de trois attributs pour lesquels on peut trouver des triplets de valeurs qui n'apparaissent qu'une seule fois dans A. À titre d'exemple, un seul individu de A vérifie (75015, 1972, M) et un seul individu de A vérifie (75019, 1977, M).

Comme c'est aussi un quasi-identifiant pour le tableau B, les valeurs uniques du quasi-identifiant qui figurent aussi dans B permettent d'identifier des individus par leur nom et prénom.

Si l'on sait que Michel Martin du tableau B figure dans le tableau A, on peut déduire que Pseudo2 désigne l'individu Michel Martin et qu'il souffre de diabète sucré. En revanche, on ne peut pas dire de quoi souffre Jean Martin, même si l'on sait qu'il figure dans le tableau A, car il pourrait tout aussi bien être Pseudo1 que Pseudo10 ou Pseudo11. De même on ne peut rien conclure sur l'individu qui pourrait être sous Pseudo 9 car la valeur de son quasi-identifiant dans A ne figure pas dans B.

Il est à noter que le croisement de la base de données contenant des informations sensibles ne nécessite pas toujours le recours à une autre base. On peut croiser la base de données avec elle-même si on dispose des versions successives de la base de données à divers instants. Par exemple, supposons qu'on connaisse la base de données A et sa mise à jour un mois plus tard, dans laquelle trois lignes ont été supprimées, suite à la sortie d'hôpital des individus Pseudo7, Pseudo8 et Pseudo9. Si on sait par ailleurs (connaissance de l'attaquant) que Joëlle Leclerc a été hospitalisée et est sortie le mois suivant, on peut inférer qu'elle souffrait d'hémoglobinopathie.

Pseudo	CP_adr	CP_naiss	Année_ naiss	Genre	Affections
Pseudo1	75016	50400	1968	M	Psychose maniaco-dépressive
Pseudo2	75015	50400	1972	M	Diabète sucré
Pseudo3	69100	69100	1945	F	Troubles du rythme cardiaque
Pseudo4	69100	14200	1950	F	Sclérose en plaque
Pseudo5	75016	04100	1968	F	Rien
Pseudo6	75012	06100	1964	M	Polyarthrite rhumatoïde
Pseudo7	75013	06100	1964	F	Hémoglobinopathie
Pseudo8	75019	87200	1977	M	Sarcoïdose
Pseudo9	75018	90500	1976	M	Lymphome de Hodgkin
Pseudo10	75016	75012	1968	M	Parkinson
Pseudo11	75016	21000	1968	M	Sarcoïdose

FIGURE 4.1 – Tableau A : Maladies avec genre, code postal adresse, code postal et année de naissance.

Nom	Prénom	Genre	CP_adr	Année_ naiss
Martin	Jean	M	75016	1968
Martin	Michel	M	75015	1972
Dupond	Jeanne	F	69100	1945
Durand	Marie	F	69100	1950
Girard	Fernand	M	69100	1928
Lambert	Julie	F	75016	1968
Leclerc	Pierre	M	75012	1964
Leclerc	Joëlle	F	75013	1964

FIGURE 4.2 – Tableau B : Base de données "publique" avec nom, prénom, genre, code postal adresse, code postal et année de naissance.

ID	Dept_adr	Année_naiss	Genre	Affections	
Pseudo1	75	]1960-1970]	M	Psychose Maniaco-dépressive	
Pseudo2	75	]1970-1980]	M	Diabète sucré	
Pseudo3	69	]1940-1950]	F	Troubles du rythme cardiaque	
Pseudo 4	69	]1940-1950]	F	Sclérose en plaque	
Pseudo 5	75	]1960-1970]	F	Rien	
Pseudo 6	75	]1960-1970]	M	Polyarthrite rhumatoïde	
Pseudo 7	75	]1960-1970]	F	Hémoglobinopathie	
Pseudo 8	75	]1970-1980]	M	Sarcoïdose	
Pseudo 9	75	]1970-1980]	M	Lymphome de Hodgkin	
Pseudo10	75	]1960-1970]	M	Parkinson	
Pseudo11	75	]1960-1970]	M	Sarcoïdose	

FIGURE 4.3 – Tableau AA "2-anonymisé"

4.2 Des techniques d'anonymisation

4.2.1 Anonymiser des données tabulaires

Nous continuons avec cet exemple pour illustrer quelques techniques d'anonymisation pour les données relationnelles ou tabulaires. On trouvera une présentation plus détaillée et plus complète de ces techniques dans l'article de Claude Castelluccia et Benjamin Nguyen [1].

Un critère de *k-anonymat* a été introduit au début des années 2000 [2]. Étant donné une table et un quasi-identifiant pour cette table, ce critère garantit que chaque valeur du quasi-identifiant apparait au moins k fois (on obtient des classes d'équivalence de k données). Avec cette contrainte, le n-uplet de valeurs du quasi-identifiant étant confondu avec au moins k-1 autres n-uplets, la probabilité de retrouver la ligne correcte dans le tableau est inférieure ou égale à 1/k. C'est alors au responsable du traitement de choisir la valeur de k, compte tenu du risque de ré-identification qu'il est prêt à accepter.

Pour anonymiser les données de A, plusieurs techniques sont possibles, telles que la suppression ou la généralisation. Par exemple, ici, on ne supprimera pas l'attribut Genre car jugé important pour l'étude des maladies, mais on pourra supprimer CP_naissance, car jugé non pertinent pour l'exploitation qu'on veut faire du tableau A. Mais cela n'est pas suffisant. On généralisera alors par exemple le code postal de l'adresse en retenant seulement le département, ainsi que l'année de naissance en se limitant à des périodes de 10 années. On obtient alors le tableau AA de la Figure 4.3. Ce tableau est 2-anonymisé par rapport au quasi-identifiant {Dept_adr, Année_naiss, Genre}, car pour chaque triplet de valeurs (Dept_adr, Année_naiss, Genre), il existe au moins deux entrées dans la table lui correspondant. Ainsi, on peut maintenant seulement dire que Michel Martin qui appartient au groupe (75,]1970-1980], M) a une chance sur 3 de souffrir de diabète sucré.

Toutefois, le critère de k-anonymat ne met pas à l'abri des attaques d'homogénéité, dans le cas où les données regroupées dans les classes d'équivalence ont toutes la même valeur sensible. Diverses extensions du modèle de k-anonymat ont été introduites pour y remédier, telles que la l-diversité , qui assure que pour chaque valeur de quasi-identifiant correspondant à k données, il y a au moins l valeurs représentatives

1. Benjamin NGUYEN et Claude CASTELLUCCIA. « Techniques d'anonymisation tabulaire : concepts et mise en oeuvre ». In : *1024 : Bulletin de la Société Informatique de France* 15 (avr. 2020), p. 23-41. URL : https://hal.archives-ouvertes.fr/hal-02570847.

2. Latanya SWEENEY. « Achieving k-Anonymity Privacy Protection Using Generalization and Suppression ». In : *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (2002), p. 571-588. URL : https://doi.org/10.1142/S021848850200165X.

pour les données sensibles. La base AA de l'exemple est 2-anonyme et 2-diverse, car on trouve toujours au moins deux maladies différentes au sein d'un groupe d'individus ayant la même valeur pour le quasi-identifiant.

Supposons maintenant que Pseudo3 soit en fait atteint de sclérose en plaque (et non de troubles du rythme cardiaque). La table AA serait encore 2-anonyme, mais ne serait plus 2-diverse car il y aurait un groupe où toutes les données auraient même valeur sensible : le groupe (69,]1940-1950], F). On saurait alors de façon sûre que Marie Durand qui appartient à ce groupe est atteinte de sclérose en plaques, alors qu'auparavant on ne pouvait le prédire qu'avec une probabilité de 1/2.

Outre les critères de k-anonymat et de l-diversité, d'autres méthodes d'anonymisation ont également été proposées dans la littérature scientifique comme la t-proximité ou la δ-divulgation [3].

La ***confidentialité différentielle*** (*Differential privacy*, DP), introduite en 2006 [4], caractérise une opération (ou exécution d'un algorithme) probabiliste sur des données, qui présente certaines garanties de confidentialité. Ce n'est pas une technique d'anonymisation à proprement parler, mais elle se combine à des techniques d'anonymisation et permet de quantifier un risque de ré-identification. Le principe est le suivant. Prenons un algorithme *ALG* qu'on exécute sur un ensemble de données $D1$. Considérons une donnée d n'appartenant pas à $D1$. La DP garantit qu'il sera très difficile de savoir en regardant le résultat de l'exécution de *ALG* sur un jeu de données D, s'il a été exécuté sur $D = D1$ ou sur $D = D1 \cup \{d\}$. Autrement dit, le résultat obtenu par *ALG* ne change pas beaucoup qu'on utilise ou non la donnée d en entrée, en plus des données de $D1$. La garantie proposée par la $\varepsilon-\delta$-DP est que la probabilité d'observer une valeur plutôt qu'une autre ne doit pas être sensiblement différente selon qu'un individu est présent ou pas.

Si une littérature abondante existe depuis plusieurs dizaines d'années concernant l'anonymisation des données tabulaires, la question de l'anonymisation, fondée scientifiquement, d'autres types de données (textes libres, documents audio et vidéo, etc.) est généralement moins avancée. Nous évoquons plus brièvement les travaux dans ces domaines.

4.2.2 Anonymisation d'autres types de données

4.2.2.1 Données textuelles

Dans de nombreux domaines, par exemple en médecine ou en sociologie, les chercheurs recourent à des textes écrits en langage naturel qui peuvent contenir des données personnelles (comme les comptes rendus cliniques ou les échanges épistolaires) et doivent répondre à des exigences de protection de la vie privée, en particulier, en Europe, à celles posées par le RGPD. D'où la nécessité de développer des techniques d'anonymisation de données textuelles en faisant appel aux ressources du traitement automatique du langage naturel (TALN)

Dans ce domaine, les tentatives d'anonymisation passent par le remplacement des identifiants textuels ; on parle alors de désidentification ou de pseudonymisation (cf. 2.4) des entités nommées, car on procède au retrait ou au masquage de toute information relevant de catégories prédéfinies correspondant aux personnes, aux lieux et aux organisations.

Un article de Meystre et co-auteurs [5] présente un état de l'art sur diverses méthodes de désidentification pour des textes relevant du domaine médical.

---

3. Nguyen et Castelluccia, « Techniques d'anonymisation tabulaire : concepts et mise en oeuvre ».

4. Cynthia Dwork. « Differential Privacy ». In : *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*. 2006, p. 1-12. url : https://doi.org/10.1007/11787006%5C_1.

5. Stephane Meystre et al. « Automatic de-identification of textual documents in the electronic health record: A review of recent research ». In : *BMC medical research methodology* 10 (août 2010), p. 70. doi : 10.1186/1471-2288-10-70.

Les grandes étapes classiques d'un processus de désidentification de données textuelles sont les suivantes :

1) déterminer les types d'entités nommées du texte et, parmi celles-ci, celles qui permettraient de relier une donnée à une personne, et qui sont donc à remplacer ou à masquer. Cette catégorie d'entités comprend les noms des personnes, les adresses (y inclus les adresses email et IP), les codes postaux, les villes, et, plus généralement, toutes les informations identifiantes telles que les numéros de téléphone ou les numéros de sécurité sociale et les dates de naissance [6] ;
2) reconnaître les entités nommées (*Named Entity Recognition* – NER) et les relations entre ces entités. Cette étape est difficile et fait l'objet de recherches actives. Les techniques de NER peuvent être à base de règles et de connaissances, et plus récemment d'apprentissage (Support Vector Machines, champs aléatoires conditionnels [7] ou réseaux neuronaux [8]) ;
3) désidentifier. Dans ce but, on peut soit retirer brutalement les identifiants directs, soit les remplacer par des noms de catégories ou des données génériques construite avec des chaînes de caractères constantes, des valeurs aléatoires ou encore des pseudonymes.

Une étude plus récente de Chevrier *et al.* [9] fait le point sur l'utilisation de l'anonymisation et de la désidentification dans le domaine de la littérature biomédicale et ses limites. Les auteurs alertent sur la nécessité pour les chercheurs de bien définir les termes qu'ils emploient, en renvoyant au besoin aux définitions données ci-dessus qui sont celles du RGPD et du HIPAA. La nécessité de bien définir ces termes est d'autant plus importante que leur mauvaise compréhension et leur mauvais usage augmentent les risques de ré-identification.

Sur ce dernier point, des travaux sur la possibilité de ré-identifier des documents textuels médicaux en français, qui avaient été automatiquement désidentifiés, ont été menés ces dernières années, en tenant compte des compétences de l'attaquant. Il ressort de cette étude [10] que la capacité à identifier des données de santé protégées est reliée à la connaissance qu'un attaquant a des documents (par exemple un soignant) et/ou de la méthode de désidentification utilisée. Elle montre que, sans connaissance spécifique ni sur le corpus ni sur la méthode, les attaquants n'ont pas pu ré-identifier de patients lorsqu'ils n'avaient pas accès à la base de patients de l'hôpital. En revanche, en ayant accès au système d'information, des patients avaient été ré-identifiés par recoupement de l'information trouvée dans plusieurs documents relatifs à ces patients et en exploitant une connaissance médicale des codes médicaux.

La communauté de chercheurs en TAL dispose de corpus désidentifiés : MIMIC-II (*Multiparameter Intelligent Monitoring in Intensive Care II*) et MIMIC-III [11]. Les données sont désidentifiées, au sens où les éléments identifiants listés dans HIPAA, les PHI, ont été retirés, et les champs de texte libre, tels que les rapports de diagnostics et les notes des médecins ont été retirés. MIMIC contient des dossiers de patients en soins intensifs, disponibles à des fins de recherche, mais nécessite toutefois un contrat

6. Aux États-Unis, dans le domaine médical, il s'agit des identifiants personnels de santé, PHI (cf. 2.4).

7. H. Yang et J.M. Garibaldi. « Automatic detection of protected health information from clinic narratives ». In : *Journal of Biomedical Informatics* 58 (déc. 2015). © 2015 Elsevier Inc. Made available under a Creative Commons license. http://creativecommons.org/licenses/by-nc-nd/4.0/, S30-S38. URL : https://eprints.whiterose.ac.uk/108935/.

8. Tanbir Ahmed, Md Momin Al Aziz et Noman Mohammed. « De-identification of electronic health record using neural network ». In : *Scientific Reports* 10.1 (oct. 2020). ISSN : 2045-2322. URL : https://doi.org/10.1038/s41598-020-75544-1.

9. Raphaël Chevrier et al. « Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review ». eng. In : *Journal of medical Internet research* 21.5 (mai 2019). Publisher: JMIR Publications, e13484-e13484. ISSN : 1438-8871. DOI : 10.2196/13484. URL : https://pubmed.ncbi.nlm.nih.gov/31152528.

10. Cyril Grouin, Nicolas Griffon et Aurélie Névéol. « Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? » In : *Proceedings of the Sixth Int. Workshop on Health Text Mining and Information Analysis*. Jan. 2015, p. 31-39. DOI : 10.18653/v1/W15-2604.

11. https://mimic.physionet.org

de mise à disposition contraignant. La base i2b2 (*Informatics for Integrating Biology & the Bedside* [12]) propose aussi des jeux de données sur des dossiers de sortie de patients désidentifiés pour des challenges internationaux de recherche sur le traitement automatique de la langue.

Chevrier et co-auteurs [13] concluent leur étude par plusieurs recommandations que nous reprenons à notre compte :

Recommandation pour les Scientifiques - 4

— Dans leurs publications, les chercheurs doivent préciser les opérations effectuées (pseudonymisation, désidentification ou anonymisation), en se référant à la terminologie existante (RGPD et HIPPA).

Recommandation pour les Scientifiques - 5

— Mettre en place un groupe pluridisciplinaire avec des compétences techniques, éthiques et juridiques, pour élaborer un guide pour aider les chercheurs dans leur démarche d'anonymisation ou de désidentification de données textuelles.

Recommandation pour les institutions de recherche et d'enseignement - 3

— L'information et la formation des chercheurs sur les questions d'anonymisation, de pseudonymisation ou de désidentification des données textuelles doivent être améliorées.

4.2.2.2 Données audio

Le sujet de l'anonymisation de données audio est encore balbutiant au niveau scientifique. À ce jour, ce qu'on qualifiait d'anonymisation audio consistait plutôt à transformer la voix des personnes concernées, le plus souvent à l'aide de dispositifs assez rudimentaires. Aujourd'hui, on peut distinguer deux grandes approches complémentaires pour l'anonymisation de données audio.

La première approche est basée sur la perception humaine. Elle propose d'aller plus loin qu'un simple floutage vocal et de rendre méconnaissable l'identité vocale d'une personne de manière extrêmement réaliste (modification du timbre, de la hauteur de la voix ou encore de l'intonation). Certains travaux, en particulier ceux menés à l'Ircam (Institut de Recherche et Coordination Acoustique/Musique) s'inscrivent dans cette direction.

La seconde approche consiste à se baser non plus sur la perception humaine mais sur la perception « machine ». Il s'agit ainsi dans ce cas, de s'assurer que la voix transformée ne peut plus permettre l'authentification de la personne par un système de reconnaissance automatique du locuteur. En effet, perception humaine et perception machine ne répondent pas aux mêmes mécanismes, si bien que l'impossibilité d'authentifier la voix transformée d'une personne par un humain ne signifie pas nécessairement qu'il en est de même pour un système de reconnaissance du locuteur (et vice-versa).

Un article récent [14] fait le point sur la nécessité de protéger la voix et l'identité du locuteur, propose un état de l'art sur les techniques existantes et suggère des pistes. D'autres travaux développés dans le cadre de challenges i2b2/ngrid peuvent être consultés [15] [16].

12. https://www.i2b2.org
13. ibidem
14. Andreas Nautsch et al. « Preserving privacy in speaker and speech characterisation ». In : *Computer Speech & Language* 58 (2019), p. 441-480. issn : 0885-2308. doi : https://doi.org/10.1016/j.csl.2019.06.001. url : https://www.sciencedirect.com/science/article/pii/S0885230818303875.
15. Amber Stubbs, Michele Filannino et zlem Uzuner. « De-Identification of Psychiatric Intake Records ». In : *J. of Biomedical Informatics* 75.S (nov. 2017), S4-S18. issn : 1532-0464.
16. Amber Stubbs, Christopher Kotfila et Özlem Uzuner. « Automated Systems for the De-Identification of Longitudinal Clinical Narratives ». In : *J. of Biomedical Informatics* 58.S (déc. 2015), S11-S19. issn : 1532-0464. url : https://doi.org/10.1016/j.jbi.2015.06.007.

4.2.2.3 Données images et vidéo

Comme nous l'avons vu de manière générale, si une information sensible n'est pas utile à la recherche envisagée, la manière la plus sûre de l'anonymiser est de la supprimer. Ainsi, si l'on cherche à anonymiser une plaque d'immatriculation ou un badge car ils ne sont pas pertinents pour la tâche à effectuer, on pourra remplacer la zone à anonymiser par une partie noire. L'information contenue dans cette zone de l'image est alors détruite et un attaquant, aussi bien outillé soit-il, ne pourra pas la reconstruire. C'est valable aussi pour des visages de personnes dans une vidéo. Si ce n'est pas gênant pour la tâche à accomplir, il faut détruire cette information, typiquement l'enlever.

De façon similaire à ce qui est observé pour les données audio, les travaux de recherche menés peuvent également être classés selon deux grandes familles d'approches basées sur la perception humaine ou la perception machine. Ainsi, dans certains travaux récents [17] [18] basés sur l'utilisation de réseaux génératifs adversaires (*Generative Adversarial Networks*, GANs), des visages de synthèse sont substitués rendant ainsi complexe la ré-identification d'un individu par un être humain. Pour la perception machine, il s'agit de se baser sur le manque de robustesse des modèles issus de l'apprentissage automatique. Ainsi certaines solutions [19] proposent d'utiliser des méthodes de type *One pixel attack* [20] pour générer des images qui ne puissent pas être utilisées dans des systèmes de reconnaissance de visage tout en restant quasi identiques pour un utilisateur humain.

Mais les techniques d'anonymisation d'un individu dans une photo ou une vidéo évoluent rapidement. À ce jour, le floutage peut être effectué grâce à de nombreuses applications mobiles, comme par exemple Facepixeliser, Photo Blur Wallpaper Booth App ou encore ObscuraCam. Mais les résultats ne sont pas toujours robustes aux techniques de dé-floutages. Certaines de ces applications suppriment toutes les données d'identification Exif (*Exchangeable image file format*) contenues dans les fichiers, y compris les données de localisation GPS et la marque et le modèle du téléphone. Pour conclure cette partie, en général, toute approche tentant de garder une partie du signal de l'image d'origine est fortement sujette à caution, voir ainsi [21]. De tels résultats ne permettent pas toujours (aujourd'hui) de formellement identifier la personne mais avec plus de données ou en perfectionnant ce type d'algorithme, ceci sera probablement faisable dans quelques années.

4.2.2.4 Données de mobilité

Les données séquentielles, par exemple de type spatio-temporelles comme les données de mobilité, ont également donné lieu à des travaux de recherche sur les techniques d'anonymisation ces dernières années, que nous évoquons plus brièvement [22]. On peut souligner l'utilité des données de transport pour les sciences sociales et les pouvoirs publics. Ces données sont par nature difficiles à anonymiser car elles sont très identifiantes [23]. En effet, des études ont montré que la connaissance de 3 ou 4 points spatiaux-

17. https://www.technologyreview.com/f/614323/ai-deepfakes-anonymizes-faces-in-videos-photos

18. Håkon Hukkelås, Rudolf Mester et Frank Lindseth. *DeepPrivacy: A Generative Adversarial Network for Face Anonymization*. 2019. arXiv : 1909.04538 [cs.CV].

19. https://findbiometrics.com/new-d-id-solution-removes-identifying-features-peoples-faces-082102/, article du 21/08/2019.

20. Jiawei Su, Danilo Vasconcellos Vargas et Kouichi Sakurai. « One Pixel Attack for Fooling Deep Neural Networks ». In : *IEEE Transactions on Evolutionary Computation* 23.5 (oct. 2019), p. 828-841. issn : 1941-0026. url : http://dx.doi.org/10.1109/TEVC.2019.2890858.

21. Sumit Rajgure et al. « Reconstructing Obfuscated Human Faces with Conditional Adversarial Network ». In : *Machine Learning and Information Processing*. Sous la dir. de Debabala Swain, Prasant Kumar Pattnaik et Pradeep K. Gupta. Singapore : Springer Singapore, 2020, p. 95-104. isbn : 978-981-15-1884-3.

22. Gergely Ács, Gergely Biczók et Claude Castelluccia. « Privacy-Preserving Release of Spatio-Temporal Density ». In : *Handbook of Mobile Data Privacy*. Sous la dir. d'Aris Gkoulalas-Divanis et Claudio Bettini. Springer, 2018, p. 307-335.

23. Arturs Lavrenovs et Karlis Podins. « Privacy violations in Riga open data public transport system ». In : nov. 2016, p. 1-6. doi : 10.1109/AIEEE.2016.7821808.

temporels d'une trajectoire suffisait pour ré-identifier, avec une probabilité élevée, une personne dans une population de plusieurs millions d'individus [24]. Différents types d'anonymisation ont été proposés dans la littérature. Certaines solutions proposent de publier uniquement des statistiques sur les différentes trajectoires, comme leur longueur moyenne ou les endroits les plus souvent visités. D'autres approches proposent de publier des données synthétiques, c'est à dire des trajectoires générées artificiellement à partir des caractéristiques statistiques des vraies trajectoires [25]. Finalement, d'autres solutions proposent de modifier les trajectoires avant de les publier, par exemple, en groupant les trajectoires similaires [26] ou en y ajoutant du bruit [27]. Pour une synthèse de ce type de technique voir [28].

4.3 Limites des techniques d'anonymisation

Il faut désormais penser globalement toute « l'économie » de la donnée et de son cycle de vie, en prenant en compte l'évolution des contextes des connaissances scientifiques, de la disponibilité de nouvelles bases de données, des déclassifications d'information, etc. Dans ce contexte, nous faisons les constats suivants.

- Les méthodes de type k-anonymat sont adaptées aux données relationnelles (tabulaires), tandis que la contrainte de confidentialité différentielle peut se généraliser plus facilement.
- Les approches de confidentialité différentielle permettent (du moins théoriquement) d'offrir des garanties qui autrement permettraient des attaques par corrélation de divers jeux de données « anonymes » produits.
- La technique de k-anonymat repose sur la notion de quasi-identifiant. Or la recherche exhaustive de tous les identifiants d'une table est exponentielle dans le nombre d'attributs (colonnes) de la relation et linéaire dans la taille de la table, dans le pire des cas, donc très coûteuse.
- Au contraire de la confidentialité différentielle, la k-anonymisation n'offre pas de garanties formelles. Ainsi, le risque de ré-identification pour une base k-anonymisée dépend des connaissances annexes de l'attaquant. Si on détermine un quasi-identifiant pour une base de données contenant des informations sensibles, le risque dépend de la probabilité que l'attaquant dispose d'une autre base avec des informations identifiantes et ce même quasi-identifiant. Or beaucoup de combinaisons de descripteurs peuvent être utilisées pour ré-identifier un individu et de plus en plus de sources de données contiennent des informations sur les individus. Dans un article publié en 2019, les auteurs affirment que 15 attributs démographiques rendraient 99.98% des individus du Massachusetts uniques [29] et donc que la connaissance de ces attributs permettrait de ré-identifier un individu.
- La confidentialité différentielle (DP) ne fait pas d'hypothèse sur le niveau de connaissances de l'attaquant. L'inconvénient toutefois est que deux instances d'anonymisation peuvent aboutir à deux résultats très différents. Si la confidentialité différentielle permet en théorie de quantifier exactement le coût de la publication d'une donnée, il reste le problème du choix du budget, et du nombre de publications qu'on s'autorise. Enfin, notons que généralement, implémenter correctement la DP est difficile, ce qui constitue un obstacle à son utilisation en pratique.

24. Yves-Alexandre de Montjoye et al. « Unique in the Crowd: The privacy bounds of human mobility ». In : *Scientific Reports, Nature* (mars 2013). URL : https://www.nature.com/articles/srep01376.

25. Gergely Acs et al. *Differentially Private Mixture of Generative Neural Networks*. 2018. arXiv : 1709.04514 [cs.LG].

26. Osman Abul, Francesco Bonchi et Mirco Nanni. « Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases ». In : *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. ICDE '08. USA : IEEE Computer Society, 2008, p. 376-385. ISBN : 9781424418367. URL : https://doi.org/10.1109/ICDE.2008.4497446.

27. Ács, Biczók et Castelluccia, « Privacy-Preserving Release of Spatio-Temporal Density ».

28. Marco Fiore et al. « Privacy in trajectory micro-data publishing: a survey ». In : *Transactions on Data Privacy* 13 (2020), p. 91-149. URL : https://hal.inria.fr/hal-02968279.

29. Luc Rocher, Julien M. Hendrickx et Yves-Alexandre de Montjoye. « Estimating the success of re-identifications in incomplete datasets using generative models ». en. In : *Nature Communications* 10.1 (déc. 2019). ISSN : 2041-1723. DOI : 10.1038/s41467-019-10933-3. URL : http://www.nature.com/articles/s41467-019-10933-3 (visité le 10/01/2022).

— On ne peut pas garantir l'anonymat de la base transformée mais on peut évaluer le risque de ré-identification (garantie probabiliste). La confidentialité différentielle est un moyen d'évaluer ce risque. Le paramétrage de ce risque n'est pas simple, en particulier lorsque l'ensemble des données sensibles est grand.

Donc en l'état actuel des connaissances, il n'existe pas de solution d'anonymisation générique, préservant suffisamment de contenu pour être utile, qui s'appliquerait à tous les types d'applications et de données.

Ce constat amène à conduire des analyses d'impact relatives à la protection des données (AIPD) que nous présentons section 6.1.3.

4.4 Evaluation et certification des techniques d'anonymisation

4.4.1 Evaluation des techniques d'anonymisation

Compte tenu des limites des techniques actuelles d'anonymisation, évaluer la difficulté de ré-identification des résultats obtenus par une technique donnée constitue un enjeu majeur.

En pratique, les autorités de protection des données se basent depuis avril 2014 sur l'avis sur les techniques d'anonymisation produit par le G29 (rassemblement des autorités de protection des données européennes, depuis remplacé par le Comité européen sur la protection des données, CEPD) [30]. Celui-ci propose un tour d'horizon des techniques d'anonymisation les plus usuelles et établit trois critères permettant d'indiquer qu'un jeu de données est anonyme. Nous reprenons leur définition et les exemples proposés par la CNIL [31] pour les illustrer :

— **Individualisation :** Est-il toujours possible d'isoler un individu dans le jeu de données ?
Exemple : une base de données de CV où seuls les nom et prénoms d'une personne auront été remplacés par un numéro (qui ne correspond qu'à elle) permet d'individualiser cette personne. Dans ce cas, cette base de données est considérée comme pseudonymisée et non comme anonymisée.

— **Corrélation :** Est-il possible de relier entre eux des enregistrements relatifs à un individu ? *Exemple : une base de données cartographique renseignant les adresses de domiciles de particuliers ne peut être considérée comme anonyme si d'autres bases de données, existantes par ailleurs, contiennent ces mêmes adresses avec d'autres données permettant d'identifier les individus.*

— **Inférence :** Peut-on déduire des informations concernant un individu ?
Exemple : si un jeu de données supposément anonyme contient des informations sur le montant des impôts de personnes ayant répondu à un questionnaire, que tous les hommes ayant entre 20 et 25 ans qui ont répondu sont non imposables, il sera possible de déduire, si on sait que M. X, homme âgé de 24 ans, a répondu au questionnaire, que ce dernier est non imposable.

Dans le cas où le respect de ces trois critères est démontré, le jeu de données peut alors être considéré anonyme au regard de la réglementation sur la protection des données. Si cela n'est pas le cas, il est alors conseillé dans l'avis du G29 de recourir à une approche par l'étude des risques. Il n'existe toutefois pas encore de consensus en Europe sur la façon de mener une telle étude. Il est à noter que s'assurer du respect de ces trois critères est loin d'être évident. Des chercheurs ont par exemple récemment montré comment on peut inférer des attributs privés dans le système Diffix évalué comme satisfaisant les exigences du G29 [32].

De nombreux travaux, principalement d'origine nord-américaine proposent des éléments méthodolo-

30. https://www.cnil.fr/fr/le-g29-publie-un-avis-sur-les-techniques-danonymisation

31. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

32. Andrea Gadotti et al. « When the Signal is in the Noise: Exploiting Diffix's Sticky Noise ». en. In : *Proceedings of the 28th USENIX Security Symposium* (août 2019), p. 19. URL : https://www.usenix.org/conference/usenixsecurity19/presentation/gadotti.

giques pour effectuer des ***mesures sur le risque de ré-identification*** au sein d'un jeu de données [33]. Trois grands types de risques sont en particulier identifiés dans la littérature :

— Risque du procureur (*prosecutor risk*) : l'attaquant souhaite savoir si un individu en particulier est présent dans un jeu de données afin d'accéder aux informations le concernant. En reprenant l'exemple donné dans le tableau A de la section 4.1 (Figure 4.1), il s'agirait ainsi d'être en mesure de prouver que Michel Martin est dans le tableau A sous le nom de Pseudo2 et ainsi d'affirmer que celui-ci est atteint de diabète sucré.

— Risque du journaliste (*journalist risk*) : l'attaquant ne se soucie pas de savoir qui est l'individu mais veut prouver qu'il est possible d'accéder aux données d'une personne. Avec l'exemple du tableau A, il s'agirait de trouver l'identité d'au moins une personne présente dans la base, par exemple Pierre Leclerc, indépendamment des caractéristiques de celle-ci. Un exemple médiatisé est le cas de Thelma Arnold, qui a été identifiée en 2006 par le NewYork Times, comme l'utilisatrice no° 4417749 après analyse de l'ensemble des requêtes sur le web d'AOL [34].

— Risque de l'annonceur (*marketer risk*) : l'attaquant souhaite ré-identifier autant d'individus que possible dans un jeu de données. Toujours avec l'exemple donné dans le tableau A, on peut imaginer qu'un laboratoire pharmaceutique mal-intentionné souhaiterait ré-identifier autant de patients que possible dans la base afin de leur proposer des services payants adaptés à la pathologie dont ils souffrent, par exemple en croisant le tableau A avec une base de clients possédant des informations de contact (adresse postale, email, numéro de téléphone, etc.), qui pourrait être le tableau B (Figure 4.2) étendu. Le critère de succès dans ce cas est la proportion des personnes ré-identifiées dans la base.

Qui plus est, cette analyse doit être régulièrement renouvelée afin de s'assurer que des risques de ré-identification évalués comme suffisamment faibles à un instant t ne deviennent pas trop importants, par exemple à la suite d'avancées scientifique ou technologiques dans le domaine.

4.4.2 Certification et homologation des techniques d'anonymisation

En 2016 la loi pour une République numérique a octroyé à la CNIL la possibilité de *« certifier ou homologuer et publier des référentiels ou des méthodologies générales aux fins de certification de la conformité à la présente loi de processus d'anonymisation des données à caractère personnel »* [35]. Concrètement, la CNIL à mis en place un ensemble de recommandations « Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation » [36]. En février 2023, la CNIL a publié une charte d'accompagnement des professionnels pour conseiller les personnes et organismes qui mettent en œuvre ou envisagent de mettre en œuvre des traitements automatisés de données à caractère personnel [37]. Il s'agit de conseils et la CNIL précise que « l'accompagnement à la conformité décrit dans la présente charte ne saurait servir à «régulariser» des comportements en cours ou passés contraires à la réglementation. »

Recommandation générale - 1

— Bien comprendre qu'on ne sait pas certifier aujourd'hui des procédés d'anonymisation.

---

33. Khaled El Emam. *Guide to the de-identification of personal health information*. eng. Taylor & Francis, 2013. ISBN : 978-1-4665-7908-8.

34. https://www.nytimes.com/2006/08/09/technology/09aol.html

35. Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique (Article 60) ayant modifié la Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés (Article 8-i)

36. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

37. https://www.cnil.fr/sites/default/files/atoms/files/charte_accompagnement_des_professionnels.pdf

— L'expression « donnée anonymisée » ne devrait être utilisée que si le procédé d'anonymisation a été certifié.
— L'expression « donnée à faible risque de ré-identification » (DFRR) devrait être utilisée à l'issue d'une analyse d'impact relative à la protection des données (AIPD) concluant que le risque d'une possible ré-identification est maîtrisé et que les impacts potentiels sur la vie privée sont considérés comme faibles.

Recommandation pour les Scientifiques - 6

— Partant du constat que la certification et l'homologation des procédés d'anonymisation sont aujourd'hui difficiles à mettre en place, il est recommandé aux scientifiques devant anonymiser des données, par exemple à des fins de publication, de faire valider leur procédé par le Délégué à la Protection des Données (DPD) (*Data Protection Officer* en anglais – DPO) de leur établissement.

Recommandation pour le grand public - 1

— Être conscient que l'anonymisation des données personnelles ne peut pas être certifiée actuellement sans une perte significative d'utilité et qu'un risque de ré-identification existe.
— Être conscient du caractère évolutif et relatif des techniques d'anonymisation.

5 Exemples de situations

Pour aborder, de façon pratique, les enjeux d'éthique de la gestion de données personnelles, nous les illustrons dans quelques domaines majeurs : santé, génomique, décisions de justice, images et vidéos dans les réseaux sociaux et les lieux publics, et enfin éducation.

5.1 Santé

5.1.1 Définition des données de santé et discussion de l'évolution de notions clés

Comme défini par la CNIL [1], « les données à caractère personnel concernant la santé sont les données relatives à la santé physique ou mentale, passée, présente ou future, d'une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l'état de santé de cette personne ». Il peut donc s'agir de données « recueillies à l'occasion d'activités de prévention, de diagnostic, de soins ou de suivi social et médico-social » (article L1110-4-1 du Code de la Santé Publique). Il peut s'agir de résultats d'examens cliniques, d'analyses biologiques, d'imagerie statique et dynamique, de tests divers, mais également de données médico-administratives. Elles peuvent être regroupées dans le Dossier Médical Partagé (DMP) et depuis le premier trimestre 2022 dans Mon espace santé [2].

Cette définition résulte des évolutions récentes dues en particulier à l'emploi généralisé du numérique :

— Au-delà du contexte d'usage de données primaires, les projets de recherche s'effectuent souvent en utilisant des informations issues de données de santé telles que des résultats d'IRM ou d'analyses biologiques dans des contextes de recherche fondamentale. Ces données, si elles ne sont pas créées dans un contexte de soin, peuvent contenir indirectement des informations concernant la santé des sujets volontaires.

— Avec l'avènement du numérique et la possibilité de croiser des données, des situations nouvelles surgissent : même des données sans rapport *a priori* avec la santé peuvent, par croisement ou par destination (utilisées dans le parcours de soin), devenir des données de santé [3].

— En outre, avec la généralisation du marché du bien-être, le recours à des données recueillies par les personnes elles-mêmes via des applications mobiles ne peut également plus être considéré comme extérieur à la notion de données relatives à la santé : L'OMS [4] considère que « la santé est un état de bien-être physique, mental et social » et sa notion de santé intègre donc des facteurs environnementaux et sociaux (comme des données sur les habitudes de vie).

En conclusion « la notion de données de santé ne peut plus se limiter aux seules données personnelles recueillies dans le cadre du soin » [5]. Les données relatives à la santé peuvent inclure une grande variété de données : données historiques, cliniques, biologiques, d'analyses et investigations, relatives aux traitements médicaux, environnementales, socio-économiques, démographiques, comportementales (qualité de vie et habitudes), etc [6].

1. https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante
2. https://esante.gouv.fr/strategie-nationale/mon-espace-sante
3. Avis 130 du CCNE https://www.ccne-ethique.fr/sites/default/files/2021-02/avis_130.pdf
4. https://www.who.int/
5. Avis 130 du CCNE, *op cit*.
6. Avis 130 du CCNE et https://www.allistene.fr/files/2018/11/rapport_numerique_et_sante_19112018.pdf

Le contexte numérique fait des données primaires une matière première pour de nombreux acteurs aux motivations et pratiques différentes. La protection de ces données, notamment par l'anonymat, est donc devenue cruciale.

5.1.2 Spécificité de la protection des données de santé

Les données de santé font partie des données qui ont un statut de données sensibles (voir la définition en section 2.4 page 12). Elles bénéficient donc d'une protection spécifique :

A- Interdiction de traitement de ces données sauf dans une série d'exceptions (RGPD, Loi Informatique et Liberté). Il est essentiel de noter que dans ces textes, des dispositions particulières visant à faciliter la recherche scientifique sont centrales et affirmées avec force, comme rappelé par la CNIL qui a établi des référentiels [7]. Dans le cas d'une recherche scientifique utilisant des données de santé et relevant d'un questionnement en matière de santé, les démarches consistent en une simple déclaration de conformité à une méthodologie de référence, ou à défaut de conformité en tous points à ce référentiel, en une autorisation préalable de la CNIL [8]. En outre, une analyse d'impact relative à la protection des données (AIPD) doit être réalisée. Nous ne détaillons pas ici ces éléments.

B- De plus, comme le rappelle le CCNE « le Code de la Santé Publique règle le secret médical, l'hébergement des données de santé, la conformité des systèmes d'information, le partage des données, l'interdiction de procéder à une cession ou à une exploitation commerciale des données de santé ». L'anonymat et l'impossibilité d'identifier sont au centre de certaines de ces notions (secret médical, sécurité des données hébergées, ...).

5.1.3 Tensions propres aux données de santé

L'ouverture des données de santé est essentielle à la recherche. Elle est prévue dans les textes (depuis le plan national pour la science ouverte de la ministre de la recherche du 4 juillet 2018) par un partage facilité grâce aux principes FAIR – « Facile à trouver, Accessible, Interopérable, Réutilisable [9] ». Dans ce contexte, les données de santé deviennent disponibles dans de grandes collections ou entrepôts de données numériques, notamment dans les institutions publiques ou privées, en France ou à l'étranger. Néanmoins, cette tendance peut entrer en contradiction avec les obligations et les responsabilités de protection des droits des personnes pour les personnels professionnels gérant des données de santé [10].

Le statut particulier des données de santé est passé au crible dans l'avis 130 du CCNE.

Pour ces données de santé, l'ensemble des progrès du système de soin et de recherche repose grandement sur l'analyse de ces données : la nécessaire solidarité à la base de cette analyse pourrait remettre indirectement en cause la protection de ces données de santé (au moins celles du système de soin). Une balance équilibrée entre cette solidarité et la protection des données est à penser [11].

D'autre part, la relation de confiance étant centrale dans le soin, le nécessaire maintien de cette confiance rentre d'autant plus en contradiction avec un risque accru de diffusion/ duplication/ réutilisation des données : ce point demande d'être très vigilant.

---

7. `https://www.cnil.fr/fr/traitements-de-donnees-dans-le-domaine-de-la-sante-les-referentiels-pour-simplifier-vos-demarches`, note du 27 décembre 2023.

8. `https://www.cnil.fr/fr/recherche-medicale-quel-est-le-cadre-legal`, note du 22 mai 2024.

9. En anglais : Findable, Accessible, Interoperable, Re-usable.

10. « Les participants expriment de façon unanime leur inquiétude concernant la confidentialité et la sécurité des données, et assurer la protection apparaît comme une priorité. Protection signifie garantir la confidentialité, respecter le secret médical, ne pas utiliser les données de santé hors du cadre de la recherche, ne pas vendre les données, assurer la sécurité de l'hébergement des données... » (CCNE, rapport de synthèse des états généraux de la bioéthique, juin 2018).

11. « ... mais il faut trouver la juste mesure entre le droit de décider pour soi et le droit d'être protégé, entre un choix personnel et l'intérêt collectif » (CCNE, ibidem).

Il convient donc de veiller à fournir des « explications et des informations sur le cheminement des données numériques et l'exploitation des données collectées » [12].

Concernant la nécessité de maîtriser les risques de ré-identification des données de santé, nous réitérons la recommandation du groupe de travail Numérique et Santé du CCNE [13].

Recommandation pour les Pouvoirs Publics - 1

— Il importe de se donner les moyens scientifiques, techniques et de régulation pour maitriser les risques de ré-identification à partir de bases de données dont les identifiants ont été supprimés.

En même temps, tous s'accordent pour insister sur le fait qu'il serait dommageable de ralentir l'innovation en soin ou en recherche médicale apportée par les traitements des données, notamment en période de pandémie, et que donc le « risque » présenté doit être maîtrisé mais en vaut la peine : si les solutions existantes ne sont pas suffisantes pour limiter le risque, des solutions nouvelles doivent être pensées. Il convient donc d'évaluer la balance bénéfices-risques des solutions existantes, ce qui suppose une formation minimale aux enjeux d'éthique du traitement des données avec les technologies numériques, notamment pour les professionnels de santé comme indiqué dans la cinquième recommandation de l'avis 130 du CCNE :

Recommandation pour les Pouvoirs Publics - 2

— Les professionnels de santé doivent bénéficier, lors de leur formation initiale et tout au long de leur carrière, d'une formation adaptée aux technologies numériques, aux principes éthiques qui régissent le recueil et le traitement des données, aux moyens à mettre en œuvre pour les respecter, et aux risques et biais qui résultent de leur non-respect.

De plus, une évaluation périodique de la mise en œuvre effective des dispositifs juridiques et opérationnels de protection est nécessaire (cf recommandation N° 2 de l'avis 130 du CCNE que nous faisons nôtre) :

Recommandation pour les Pouvoirs Publics - 3

— Compte tenu du rythme particulièrement important des innovations scientifiques et technologiques et des évolutions qu'elles déterminent dans le recueil et l'exploitation des données relatives à la santé, il est nécessaire d'évaluer périodiquement la mise en œuvre effective des dispositifs juridiques, afin de vérifier le maintien dans le temps de l'efficacité du système de protection des données personnelles qu'ils instaurent.

A destination des chercheurs, nous rappelons et faisons nôtre la recommandation 11 de l'avis 130 du CCNE, écrite dans le contexte des données de santé, mais qui pourrait s'appliquer à d'autres types de données :

Recommandation pour les Scientifiques - 7

— En matière de recherche, l'impératif éthique doit être adapté à chaque situation particulière, de manière à établir une relation de confiance entre les personnes dont on traite les données et celles qui y accèdent et les traitent. Il est essentiel que la personne dont on traite les données soit informée des modalités pour lesquelles l'autorité de contrôle assure sa fonction de tiers de confiance.

5.1.4 Limites à l'anonymat dans le domaine de la santé

Au-delà des limites techniques de l'anonymisation des données, il est important de souligner que certaines données relevant de la santé ne peuvent pas, de par leur nature, être anonymisées facilement car leur singularité les rend identifiantes. Cela est évident avec les données génétiques (car chaque séquence d'ADN est liée de façon unique à un individu) et cela le devient maintenant avec les données d'imagerie

12. cf. avis 130 du CCNE, page 13.

13. Numérique et santé : quels enjeux éthiques pour quelles régulations ?, rapport du groupe de travail commandé par le CCNE avec le concours de la CERNA, nov. 2018, page 47, `https://www.allistene.fr/files/2018/11/rapport_numerique_et_sante_19112018.pdf`.

cérébrale (de type IRM Imagerie par résonance magnétique). En effet, on peut identifier une personne en reconstituant via les IRM de sa tête suffisamment d'information sur l'aspect extérieur de son visage pour que ce modèle soit identifié par une personne ou un système de reconnaissance faciale standard [14].

Par ailleurs, il faut noter que les personnes peuvent être amenées à divulguer leurs données de santé, volontairement ou non, par des recherches d'informations de santé via un moteur de recherche sur Internet, la participation à des blogs de santé, l'achat en ligne de médicaments, la prise de rendez-vous sur des plateformes de télé-consultations, ou encore par l'usage d'objets connectés qui fournissent des données, qui, par recoupement peuvent renseigner sur leur état de santé et leurs habitudes de vie.

5.2 Génomique et génétique

Parmi les données de santé, les données génétiques et plus récemment les données génomiques présentent des caractéristiques notables qui soulèvent de multiples interrogations. La génétique, science de l'hérédité, qui s'intéresse aux caractères transmissibles de génération en génération, a été en effet révolutionnée par l'avènement de la génomique. La génomique étudie le fonctionnement d'un organisme à l'échelle du génome entier, c'est-à-dire de toute la macromolécule d'ADN (suite de nucléotides A, C, G, T). L'ensemble des caractéristiques génétiques (génotype) d'un individu, héréditaires ou acquises, résultant de l'analyse de son génome, donne des informations uniques sur sa physiologie ou son état de santé (phénotype). Les données génomiques contiennent ainsi des informations précieuses pour le diagnostic, le soin et le pronostic d'un certain nombre de maladies comme le cancer.

5.2.1 Qualification plurielle des données génétiques

Au sens juridique, les données génétiques sont qualifiées de données à caractère personnel et relèvent dans le RGPD de la catégorie des « données sensibles ». Elles sont à ce titre expressément visées par l'article 9, qui les soumet à un régime spécifique [15]. Toutefois, la CNIL [16] en ce domaine souligne l'attention toute particulière qui doit être accordée à la protection de ces données en raison de leurs caractéristiques propres et de leur fort potentiel discriminant.

Cela tient tout d'abord à leur caractère éminemment identifiant et quasi-immuable. Les données génomiques sont en effet des données « très personnelles », particulièrement intimes, en ce qu' « elles décrivent le plus profond, le plus secret, de la dimension biologique de l'individu : son patrimoine génétique », et sont « étroitement liées à des dimensions aussi personnelles que son origine, sa filiation, son état de santé, passé ou futur, ses potentialités, sa descendance, et peut-être son avenir ».

Par ailleurs, ces données sont des données « pluripersonnelles » ou « données personnelles partagées » en raison de leur nature héréditaire et transmissible : elles peuvent révéler des informations sur et avoir des implications pour les consanguins de cet individu, y compris les générations antérieures et postérieures, voire permettre de caractériser un groupe de personnes, comme une communauté ethnique [17].

14. Schwarz et al., « Identification of Anonymous MRI Research Participants with Face-Recognition Software » ; Tonya White, Elisabet Blok et Vince D. Calhoun. « Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed ». In : *Human Brain Mapping* 43.1 (2020), p. 278-291. doi : https://doi.org/10.1002/hbm.25120.

15. V. également CEDH, Marper c. RU sur les données ADN qualifiées de Données à caractère personnel (CEDH, 4. 12. 2008)

16. CNIL, *Les données génétiques*, La documentation française, 2017, p. 23. Et la CNIL de rappeler que l'on considère usuellement l'ADN, « trace individuelle unique », « comme l'outil d'identification par excellence » (p. 29).

17. CNIL, op. cit., p. 43.

5.2.2 Tensions et perspectives sans précédent en génomique

On assiste désormais à la constitution de grandes bases de données génomiques. Ces projets, portés par différents États et opérateurs, sont tout à la fois sources d'opportunités formidables et de risques majeurs. Comme le relève la CNIL dans ses travaux, les données génétiques constituent un immense gisement d'informations à interpréter, « auxquelles il ne reste qu'à donner du sens », ce en raison notamment des progrès considérables réalisés grâce au développement de nouvelles techniques d'analyse et d'ingénierie génomique comme les méthodes de séquençage à haut débit (*New Generation Sequencing*, NGS), couplé à la capacité à traiter des grandes masses de données, la réduction des coûts de séquençage et l'augmentation des capacités de stockage dans le Cloud. Les données génomiques permettent dès lors d'informer « sur le passé génétique des individus et des populations dont ils font partie, sur l'histoire de leurs ancêtres » et livrent, « par le jeu des corrélations statistiques, des éléments d'information sur les prédispositions à telle ou telle pathologie » par l'identification de marqueurs génétiques, ce qui présente des perspectives inédites en termes de médecine prédictive. Dans le même temps, le dévoilement de ces données est susceptible de nuire grandement aux individus, car elles mettent à nu leurs vulnérabilités.

5.2.3 Une impossible anonymisation ?

L'anonymisation des données génomiques se heurte au caractère fortement identifiant du génome, comme signalé en section 5.2.1. En effet, même de petites parties du génome suffisent à identifier un génome complet. A ce titre, les SNP (*Single Nucleotide Polymorphism*) [18] jouent un rôle majeur pour l'identification d'individus ou de groupes d'individus (ethnies). Deux génomes humains quelconques varient au total sur 0,1 à 0,6% de leur séquence d'ADN, et ces variations génétiques correspondent à la diversité des individus, la sensibilité plus ou moins grande aux maladies ou à l'environnement, et la réponse aux médicaments.

Les SNP sont très fréquents, régulièrement répartis sur le génome et représentent la grande majorité des variations entre deux génomes d'une même espèce. Lin et co-auteurs ont montré en 2004 qu'en connaissant environ 80 SNP statistiquement indépendants, on peut identifier une personne [19].

Ces différentes caractéristiques interrogent la possibilité d'anonymiser des données génomiques, de façon pérenne, d'autant plus qu'on assiste à une collecte croissante de nouvelles données et à la conception de nouveaux traitements algorithmiques. De nombreux travaux ont ainsi montré que ces données génomiques font l'objet d'un risque très élevé de ré-identification par croisement avec tous types de données.

En 2013, des chercheurs ont ainsi montré que les noms de personnes peuvent être retrouvés à partir de leurs génomes par profilage de microsatellites (*short tandem repeats*) sur le chromosome Y, en croisant avec des bases de données de généalogie génétique récréative [20], mettant en évidence les effets du croisement de données.

Un article paru en 2014 [21] énumère différentes stratégies permettant de rompre la confidentialité des données génétiques par des techniques de ré-identification.

On relèvera qu'à l'issue de la publication de ces études, les séquences génomiques n'ont plus été publiées en données ouvertes (*open data*), mais limitées à des fins de recherche biomédicale. Cependant, les données de généalogie génétique ne sont pas soumises à des restrictions d'accès et soulèvent des pro-

18. Un SNP est observé lorsque deux génomes d'une même espèce diffèrent sur un segment d'ADN donné par une seule paire de bases, variation observée sur plus de 1% de la population.

19. Zhen Lin, Art B. Owen et Russ B. Altman. « Genomic Research and Human Subject Privacy ». In : *Science* 305.5681 (2004), p. 183-183. doi : 10.1126/science.1095019.

20. Melissa Gymrek et al. « Identifying Personal Genomes by Surname Inference ». In : *Science* 339.6117 (2013), p. 321-324. doi : 10.1126/science.1229566.

21. Yaniv Erlich et Arvind Narayanan. « Routes for breaching and protecting genetic privacy ». In : *Nature Reviews Genetics* 15 (2014), p. 409-421. doi : 10.1038/nrg3723.

blèmes de confidentialité. On observe en effet un véritable engouement pour les tests génétiques, avec une croissance importante de la génétique « directe aux consommateurs » (*Direct-To-Consumers*, DTC). De plus en plus d'entreprises, en effet, vendent des tests génétiques et rassemblent des données génétiques, avec comme application majeure la généalogie. Toutefois, la loi française de bioéthique de 2021 restreint les études des caractéristiques génétiques des personnes à des fins médicales, judiciaires ou de recherche scientifique (titre III, article 16). La généalogie génétique récréative est donc interdite en France, même avec le consentement de la personne concernée. La CNIL rappelle les risques liés à la fiabilité des résultats des tests et à l'absence de transparence sur l'utilisation des données personnelles sensibles recueillies. Les données récoltées par de telles entreprises peuvent faire l'objet de piratages comme `23andMe` en décembre 2023 [22]. La CNIL [23] précise que l'achat d'un test génétique sur Internet par des personnes résidant en France est passible de 3 750 € d'amende.

En 2018, Erlich et co-auteurs ont établi qu'on pouvait inférer l'identité d'une personne à partir de ses données génomiques et à l'aide de recherches familiales élargies dans des bases de données de généalogie génétique [24].

L'étude a également montré qu'on pouvait retrouver l'identité d'un participant à un projet de recherche génomique, comme par exemple le projet « 1000 génomes » démarré en 2008 et qui reposait sur la participation volontaire d'un millier de personnes appartenant à différents groupes ethniques, qui acceptaient qu'on séquence de façon anonyme leur génome et qu'il soit rendu public [25].

Plus récemment, des chercheurs ont montré comment des attaquants pouvaient exploiter de façon systématique ce service de téléversement de données génomiques offert par des services de génétique DTC, pour révéler l'identité des données génomiques stockées dans leur base, sans avoir été autorisés à accéder à ces données [26] [27].

Ces études mettent en évidence l'urgence de reconnaître les données génomiques concernant le génotype comme des informations fortement identifiantes et soulignent la nécessité d'informer les utilisateurs de services génétiques DTC de la fragilité de leur anonymat [28].

Il est difficile voire impossible d'anonymiser les données génomiques. En conséquence, il faut envisager un accès confidentiel aux données, éventuellement en utilisant des techniques de chiffrement, des techniques de confidentialité différentielle, des méthodes de chiffrement homomorphes, voire une combinaison de ces méthodes [29]. Ces dernières approches ouvrent des perspectives intéressantes qui se heurtent, pour l'instant, à l'ineffectivité calculatoire des techniques homomorphes. Voir par ailleurs le chapitre 6 ci-dessous.

Les particularités des données génomiques et leur caractère potentiellement fortement discriminant appellent à soumettre leur traitement à un cadre particulièrement strict. Nous reprenons certaines des re-

22. `https://www.bbc.com/news/technology-67624182`

23. `https://www.cnil.fr/fr/tests-genetiques-sur-internet-la-cnil-appelle-la-vigilance`

24. Yaniv Erlich et al. « Identity inference of genomic data using long-range familial searches ». In : *Science* 362.6415 (2018), p. 690-694. doi : `10.1126/science.aau4832`.

25. 1000 Genomes project: `https://www.internationalgenome.org/`

26. Michael D. Edge et Graham Coop. « Attacks on genetic privacy via uploads to genealogical databases ». In : *bioRxiv* (2019). doi : `10.1101/798272`.

27. Peter Ney, Luis Ceze et Tadayoshi Kohno. « Genotype Extraction and False Relative Attacks: Security Risks to Third-Party Genetic Genealogy Services Beyond Identity Inference ». In : *Network and Distributed System Security Symposium*. Jan. 2020. doi : `10.14722/ndss.2020.23049`.

28. C'est par exemple ce qu'a fait GEDMatch, à la suite de l'affaire du tueur du Golden State, tueur en série américain qui a sévi dans les années 1970 et 1980, et qui a pu être identifié en 2018, par croisement, dans la base de données généalogique GEDMatch, d'un échantillon de son ADN avec celui d'un membre de sa famille.

29. Jean Louis Raisaro et al. « Protecting Privacy and Security of Genomic Data in I2b2 with Homomorphic Encryption and Differential Privacy ». In : *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15.5 (sept. 2018), p. 1413-1426. issn : 1545-5963. doi : `10.1109/TCBB.2018.2854782`.

commandations retenues par la CNIL dans son étude [30] en les étendant :

Recommandation pour les Scientifiques - 8

— Informer les personnes dont on étudie les données génomiques de l'impact de cette étude à la fois pour elles mais aussi pour leur famille, leurs ancêtres et descendants, et plus généralement leur parentèle.

Recommandation pour les Scientifiques - 9

— Lorsqu'ils engagent des projets de recherche avec des données génomiques, les chefs de projets doivent expliciter les spécificités du consentement éclairé dans ce contexte.

Recommandation pour les Scientifiques - 10

— Le consentement doit être adossé à des garanties relatives à la qualité scientifique des études réalisées avec les données génomiques, à leur caractère respectueux des personnes et aux mesures de gouvernance des bio-banques, en lien avec les comités d'éthique opérationnels locaux.

Recommandation pour les Pouvoirs Publics - 4

— Former les citoyens et en particulier les décideurs pour qu'ils comprennent les implications, en termes de confidentialité, du télé-versement de leurs données génétiques dans une base de données de généalogie génétique, surtout si elle propose un service de génétique « direct aux consommateurs ».

Comme pour les données de santé en général, mais plus particulièrement à cause de la difficulté d'anonymiser les données génomiques, nous rappelons et faisons nôtre dans le contexte de la génomique la recommandation 11 de CCNE avis 130 pour les chercheurs déjà rappelée en conclusion de la section 5.1.3.

5.3 Décisions de justice

Le dilemme de l'anonymisation, que nous avons illustré avec les données de santé à la section 2.3.2.2, se pose de nouveau avec les données judiciaires, mais sous une forme différente. En effet, ici, ce n'est plus simplement l'intérêt de la personne du patient qui entre en tension avec celui du chercheur, mais l'intérêt des individus qui s'oppose à la fois aux intérêts des citoyens et des chercheurs : publier les décisions de justice sans les anonymiser porte atteinte à la vie privée des justiciables mais ne pas les publier, ou les publier après les avoir anonymisées, entrave à la fois le besoin des citoyens d'être informés des conclusions de la justice (besoin de transparence [31]) et le travail des chercheurs en droit (l'anonymisation des textes compliquant l'identification et la recherche des documents [32]).

Ce dilemme existe depuis que le principe de la publicité de la justice a été consacré, à la fin du XVIII^{e} siècle. Mais son importance a été longtemps limitée par la faible diffusion de l'information : les jugements étaient certes rendus publiquement, mais l'information parvenait rarement à la ville voisine où le condamné, après avoir purgé sa peine, pouvait retrouver une forme d'anonymat.

Le développement des bases de données jurisprudentielles a transformé ce dilemme, si bien que beaucoup proposent de limiter la publicité des décisions de justice pour respecter l'anonymat des personnes concernées par les décisions. Par exemple, la CNIL recommande, depuis 2001, que les éditeurs de bases de données de décisions de justice librement accessibles sur le net s'abstiennent d'y faire figurer le nom et

30. CNIL, préc., p. 118&s.

31. Emmanuel Derieux, 13 avril 2021, https://www.actu-juridique.fr/ntic-medias-presse/lanonymisation-des-decisions-de-justice-est-elle-compatible-avec-la-liberte-dexpression/

32. https://www.dalloz-actualite.fr/interview/contre-l-anonymisation-des-arrets-publies-decadence-des-references-de-jurisprudence

l'adresse des parties au procès et des témoins [33]. Notons toutefois que, tout en protégeant l'individu, cette proposition de la CNIL se fait au détriment du citoyen qui demande à être informé, en toute transparence, de l'activité du pouvoir judiciaire et des sanctions qu'il prononce. Plus récemment, le décret du 29 juin 2020 relatif à la mise à la disposition du public des décisions des juridictions judiciaires et administratives établit les conditions de publication : « Les nom et prénoms des personnes physiques mentionnées dans la décision, lorsqu'elles sont parties ou tiers, sont occultés préalablement à la mise à la disposition du public. Lorsque sa divulgation est de nature à porter atteinte à la sécurité ou au respect de la vie privée de ces personnes ou de leur entourage, est également occulté tout élément permettant d'identifier les parties, les tiers, les magistrats et les membres du greffe. Les données d'identité des magistrats et des membres du greffe ne peuvent faire l'objet d'une réutilisation ayant pour objet ou pour effet d'évaluer, d'analyser, de comparer ou de prédire leurs pratiques professionnelles réelles ou supposées. La violation de cette interdiction est punie des peines prévues aux articles 226-18,226-24 et 226-31 du code pénal, sans préjudice des mesures et sanctions prévues par la loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés. » [34] [35]. De plus, il existe une procédure de demande d'occultations complémentaires ou de levée d'occultation [36]. Ces textes utilisent le terme « occultation » qui renvoie à la désidentification (voir la section sur les données textuelles 4.2.2.1).

On assiste actuellement au développement de l'open data pour les décisions de justice [37]. Par exemple, depuis décembre 2023, les jugements rendus en matière civile, sociale et commerciale par plusieurs tribunaux judiciaires sont diffusés en *open data* sur le site de la Cour de cassation. La publication de ces jugements est rendue possible par l'utilisation d'un logiciel de pseudonymisation des décisions de justice. Cet outil est développé sous l'égide de la Cour et comporte une interface d'annotation permettant de vérifier et corriger la pseudonymisation qu'il propose [38].

Ceci nous amène à la recommandation suivante :

Question de recherche - 1

— Développer les recherches entre droit et numérique pour éclairer les évolutions possibles de la publication des données de justice.

5.4 Données d'images et de vidéos sur les réseaux sociaux et dans les lieux publics

Les réseaux sociaux comme Facebook, Instagram, YouTube, Twitter ou TikTok offrent des services de partage de photos et de vidéos largement répandus. L'utilisation de ces services s'est fortement accrue ces dernières années. Ainsi, Instagram comptait en 2018 un milliard d'utilisateurs actifs mensuels, soit 10 fois plus qu'en 2013 (source : statistica), dont 500 millions se connectaient quotidiennement. Et ce chiffre ne cesse d'augmenter puisqu'en 2022 on compte 1,39 milliards d'utilisateurs actifs [39]. Les questions d'anonymisation sur ce réseau social sont d'autant plus importantes que 71% des utilisateurs d'Instagram ont moins de 34 ans, avec un usage massif chez les adolescents. Les photos et vidéos partagées sur les réseaux

33. Open data : la protection des données comme vecteur de confiance, 29 août 2017, https://www.cnil.fr/fr/open-data-la-protection-des-donnees-comme-vecteur-de-confiance

34. Décret n° 2020-797 https://www.legifrance.gouv.fr/eli/decret/2020/6/29/JUST1933453D/jo/texte

35. Article L111-13 du Code de l'organisation judiciaire : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000038311162

36. https://www.justice.gouv.fr/documentation/open-data-decisions-justice

37. Open data des décisions de justice : où en est-on ? https://www.village-justice.com/articles/open-data-des-decisions-justice-enfin-decret-tant-attendu,35962.html

38. Cour de Cassation https://www.dalloz-actualite.fr/interview/open-data-des-decisions-des-tribunaux-judiciaires-une-nouvelle-etape-novatrice

39. https://www.oberlo.fr/blog/chiffres-instagram, consulté en juillet 2022.

sociaux communiquent des informations entre utilisateurs, dans un premier cercle d'amis tout d'abord, puis de manière virale ensuite sur tout le réseau lorsque la photo ou la vidéo est plus largement partagée. Depuis août 2016, Instagram a lancé Instagram Story, permettant aux utilisateurs de prendre des photos et des vidéos qui ne sont accessibles aux utilisateurs qu'une journée. Ce nouveau service a entraîné un partage encore plus important de photos ou de vidéos sur ce réseau social. Les contenus des photos ou vidéos partagées sur les réseaux sociaux sont variés. Ils concernent souvent des évènements vécus dans un contexte personnel et privé de l'utilisateur ; ils dévoilent alors de nombreuses informations sur sa vie quotidienne, dévoilement qui croît de jour après jour.

Les photos et vidéos des réseaux sociaux sont des données riches qui pourraient être largement utilisées aujourd'hui pour l'entraînement, par apprentissage machine, de logiciels de reconnaissance de visages, de postures, de comportements ou d'objets.

Les applications sont nombreuses dans de multiples secteurs à fort potentiel économique. Ainsi, dans le secteur du luxe, cela peut aider à détecter des imitations ou à analyser les contenus diffusés sur des marques et produits. D'un point de vue sociologique, cela permet de mesurer l'impact de la diffusion virale d'un contenu ou d'analyser des réseaux exploitant le contenu des photos ou vidéos. La reconnaissance de photos et de vidéos présente également des applications dans le domaine de la sécurité nationale : identification d'individus suspects, recherche d'individus ayant commis des actes terroristes ou plus simplement des délits. Dans cet objectif, les technologies de reconnaissance faciale connaissent aujourd'hui une grande expansion. Elles reposent sur l'exploitation de données issues de caméras placées dans des lieux publics, par exemple dans les réseaux de transports (métro dans les grandes villes) ou dans les bâtiments publics (mairies, préfectures, parlement, abords des écoles etc...). Ces techniques reçoivent aussi des applications dans le secteur multi-média, pour l'indexation de vidéos et l'accès par le contenu à des bases de données de films ou d'images.

Les technologies de traitement de photos et de vidéos soulèvent plusieurs questions d'ordre éthique : en matière de droit à l'image des personnes physiques, et bien qu'il existe des exceptions et cas particuliers, il est nécessaire de recueillir le consentement d'une personne préalablement à l'utilisation de son image. Cela apparaît d'autant plus nécessaire que la donnée d'un visage est particulièrement identifiante, plus qu'un nom (pour lequel il peut exister des homonymes) ou qu'un prénom (cf. 4.2.2.3). En toute rigueur, l'utilisation des données d'images ou de photos collectées sur des réseaux sociaux ou dans les lieux publics, et l'exploitation par des technologies permettant d'identifier les personnes, devraient requérir des techniques d'anonymisation, lorsque les individus en question n'ont pas donné leur consentement éclairé à l'usage de ces données. Toutefois, on doit pouvoir mettre en place, dans des conditions à définir, des exceptions pour la recherche, sous réserve que les résultats de l'exploitation de ces données ne conduisent pas à la diffusion d'informations identifiantes.

Dans le cadre de la régulation sur les dispositifs de reconnaissance faciale dans les lieux publics - voir en particulier le travail de réflexion mené par la CNIL en 2019[40] - l'enjeu d'éthique porte sur le compromis entre la nécessité d'assurer la sécurité des citoyens et la possibilité d'exercer une surveillance généralisée[41]. Les techniques d'authentification sont compatibles avec une anonymisation des données. Il n'en va pas de même avec les techniques d'identification d'individus qui sont, par principe, antinomiques avec une anonymisation des données puisqu'elles visent à retrouver l'identité d'une personne sans son consentement explicite. Elles posent, par conséquent, une question éthique essentielle sur la nécessité ou non de mettre en œuvre ce type de technologies dans les lieux publics. On constate d'ailleurs actuellement chez certains citoyens de nouveaux comportements d'évitement, par exemple l'utilisation de maquillages

40. https://www.cnil.fr/fr/reconnaissance-faciale-pour-un-debat-la-hauteur-des-enjeux

41. CNPEN : Comité National Pilote d'Éthique du Numérique. *Enjeux éthiques des technologies de reconnaissance faciale, posturale et comportementale.* fr. Avis CNPEN-8. Nov. 2023. URL : https://www.ccne-ethique.fr/sites/default/files/2024-02/CNPEN_AVIS8.pdf.

d'échappement à la reconnaissance faciale [42]. Si ces usages se généralisaient, cela conduirait au paradoxe selon lequel les citoyens ordinaires seraient suivis à la trace, alors que les personnes se sachant suspectes disposeraient des moyens de se camoufler.

Il existe des recherches scientifiques visant à développer d'autres moyens d'identification que la reconnaissance faciale, par exemple la reconnaissance de postures comme la démarche, la manière d'utiliser un clavier ou d'interagir avec un écran tactile etc. Les données relatives aux travaux de recherche dans ce cadre sont des données personnelles qu'il convient de traiter en tant que telles.

Point d'information : en France, on ne peut pas utiliser pour la recherche, sans autorisation préalable, les photographies et les vidéos accessibles sur les réseaux sociaux.

Point de vigilance : les chercheurs doivent prendre conscience de la difficulté technique d'anonymisation sans dégradation majeure des images et vidéos.

Recommandation pour les Scientifiques - 11

— Les chercheurs doivent veiller à ce que les données images qu'ils utilisent pour leurs recherches, en particulier celles qui sont disponibles sur les réseaux sociaux, vérifient les critères légaux requis. Il leur faut aussi s'assurer des conditions légales et respectueuses des personnes lors de la publication des résultats de recherche et de la diffusion des données permettant leur reproductibilité.

5.5 Données pédagogiques

On appelle donnée pédagogique toute information en lien avec un processus pédagogique.

Ces données concernent aussi bien les formateurs que les apprenants, ainsi que toutes les personnes accompagnant l'activité pédagogique. Elles concernent aussi le contexte dans lequel l'activité pédagogique se déroule : crèche, école maternelle, primaire, collège, lycée, université, grande école, entreprise, domicile, plateforme numérique, espace de co-working, etc.

Les exemples traditionnels sont à l'esprit de tous : les écrits et réalisations des apprenants (cahiers, copies, notes prises en cours, ...), les écrits et réalisations des formateurs (notes de support de cours, énoncés d'exercices, appréciations, notations et commentaires sur les travaux des apprenants, ...), les relevés de notes, le dossier pédagogique des apprenants, le dossier administratif des formateurs (résultat d'inspection, dossier de carrière, ...).

A ces exemples que ne renierait pas Jules Ferry s'ajoutent aujourd'hui l'ensemble des éléments apportés par les avancées scientifiques et technologiques dues en particulier au numérique, qui exposent et exploitent les données pédagogiques personnelles. Tout d'abord les plateformes numériques de support à une activité pédagogique, que ce soit sous la forme de cours en ligne dans des formes très variées, MOOCs (*Massive Open Online Course*) et leurs multiples variantes (miniMOOC, SPOC (*Small Private Online Course*), xMOOC, cMOOC, ...) ou les plateformes d'aide au suivi pédagogique des élèves en classe comme « à la maison » (e.g. Maxicours, ...) ou les vidéos à visées pédagogiques disponibles sur les plateformes de diffusion comme DailyMotion ou YouTube. Pour l'ensemble de ces plateformes, souvent remarquables de qualité dans leur offre pédagogique et dans leur ergonomie, des services adjacents sont proposés comme l'usage de réseaux sociaux numériques pour permettre les travaux et les réflexions en groupe, la possibilité de surveillance à distance de situations d'examen ou l'hébergement de contenus ou de vidéos ou encore la certification des résultats obtenus par les apprenants. D'autre part, on voit apparaître des systèmes automatisés de génération de textes comme ChatGPT [43], qui absorbent des données personnelles et pourraient modifier profondément les pratiques pédagogiques.

Les données pédagogiques révèlent des caractéristiques des personnes dont elles sont issues de manière

42. https://cvdazzle.com
43. https://fr.wikipedia.org/wiki/ChatGPT

très précise et profonde. En effet, ces profils, souvent disponibles sous forme numérique [44], sont issus de situations où les apprenants comme les formateurs s'investissent au mieux de leurs capacités et de leurs compétences, sans chercher à jouer un rôle idéalisé comme certains peuvent le faire dans des réseaux sociaux. Cette précision des données est un avantage remarquable permettant de développer des analyses de données pédagogiques (*learning analytics*). Elles incluent des données personnelles précises, par exemple les dates et heures de connexion, et profondes, par exemple au sens de ce que sait, ou ne sait pas, une personne. Les données, enregistrées systématiquement, permettent des analyses pour comprendre comment aider directement en ligne les apprenants face à leurs difficultés éventuelles : il ne sert par exemple à rien de chercher à apprendre ou utiliser le concept de multiplication si le concept d'addition n'est pas acquis préalablement. Elles peuvent également permettre à l'apprenant de s'orienter en prenant en compte des données objectivées de manière fine. Ces analyses permettent aussi aux formateurs de mieux comprendre les points sur lesquels le cours pourrait être amélioré ou comment mieux aider leurs élèves. Dans le contexte de la formation continue elles peuvent permettre aussi de comprendre les compétences des personnes en entreprise ainsi que la qualité des relations entre des personnes dans leur entreprise. Ces analyses peuvent aussi servir, ce qui est loin d'être neutre, à orienter les services des ressources humaines en quête de recrutement vers les personnes les plus adaptées aux postes recherchés. Elles peuvent également servir de faire-valoir aux apprenants en recherche de travail, de stage ou de promotion.

Les données pédagogiques sont donc d'une richesse telle qu'elles permettent, au-delà des apports qu'elles peuvent avoir pour les apprenants ou les formateurs, d'avoir une vision claire des compétences humaines d'une classe, d'une école, d'une entreprise voire d'une nation : elles constituent donc une ressource stratégique. Ces données, même agrégées, constituent donc clairement un élément de la souveraineté de chacune de ces entités.

A ce sujet, le comité d'éthique pour les données d'éducation émettait dans son avis n°2020-1 « Enjeux d'éthique des usages des données numériques d'éducation dans le contexte de la pandémie » [45] deux recommandations [46] que nous reprenons à notre compte :

Recommandation pour les Pouvoirs Publics - 5

— Définir une stratégie nationale et portée par l'Europe concernant le développement de produits numériques d'éducation respectant les valeurs fondamentales Européennes.

Recommandation pour les Pouvoirs Publics - 6

— Sensibiliser les différents acteurs (enseignants, élèves, familles, entreprises, acteurs académiques, politiques) aux enjeux de souveraineté numérique dans le domaine des données scolaires numériques.

Du point de vue de l'individu, qu'il soit en position de formateur ou d'apprenant, les données pédagogiques le concernant sont précises, pertinentes et le décrivent de manière fine et durable.

Elles peuvent révéler ainsi, du point de vue de l'anonymisation, des données sensibles, typiquement si on les croise avec d'autres informations, par exemple en combinant les absences d'un élève et les dates de fêtes religieuses [47]. Elles sont très similaires aux données de santé en ce qu'elles contribuent à décrire de manière très riche et précise les compétences, capacités, savoirs et savoir-faire, l'autonomie, la créativité, les habitudes de travail et de relation des personnes. Cependant, de façon surprenante, les données pédagogiques ne sont pas en tant que telles considérées comme des données sensibles au sens du RGPD, contrairement aux données de santé. Dans son avis n°2020-1 1e comité d'éthique pour les données

44. https://edps.europa.eu/sites/edp/files/publication/icdppc-40th_dewg-resolution_adopted_fr.pdf

45. https://www.education.gouv.fr/le-comite-d-ethique-pour-les-donnees-d-education-12146

46. Avis n°2020-1 du comité d'éthique pour les données d'éducation, recommandations 1 et 6 de la section « Garantir la souveraineté numérique en matière d'éducation ».

47. C. Zolynski et Th. Toulotte, Données traitées – Données scolaires, *in* Droit des données personnelles, Les spécificités du droit français au regard du RGPD, Dalloz, 2019.

d'éducation a formulé dans ce sens plusieurs recommandations que nous adoptons [48] :

Recommandation pour les Pouvoirs Publics - 7

— Engager une réflexion sur l'opportunité d'introduire dans le RGPD le statut de données sensibles pour les données pédagogiques, à l'instar des données de santé, à des fins de protection de la vie privée des élèves et des personnels de l'éducation.

Recommandation pour les Pouvoirs Publics - 8

— Donner un statut juridique plus protecteur pour les données d'éducation au niveau français en instituant des codes de conduite sectoriels sur les données d'éducation.

Recommandation pour les Pouvoirs Publics - 9

— Attirer l'attention des différents acteurs (enseignants, élèves, familles, entreprises, acteurs académiques, politiques) sur le caractère « spécifique » des données d'éducation et continuer à responsabiliser ces acteurs dans leur utilisation de ces données.

Recommandation pour les Pouvoirs Publics - 10

— Intensifier la formation au droit à la protection des données, en particulier de celles qui sont liées aux usages pédagogiques numériques, pour les enseignants, les élèves et leurs familles, en les illustrant avec des cas pratiques.

Recommandation pour les Pouvoirs Publics - 11

— Offrir des garanties de sécurité et de souveraineté des outils que l'État recommande ou met à la disposition des acteurs de l'éducation, et évaluer le risque que fait courir leur perte éventuelle de confidentialité.

Cette analogie entre données de santé et données pédagogiques s'étend à la manière dont elles sont collectées et mémorisées. Côté santé, nous mentionnons deux plateformes. Après le DMP, Dossier Médical Partagé, la plateforme Mon espace santé [49] permet le suivi de chaque personne au cours de sa vie et des interventions plus rapides et informées en cas de problème de santé. Le Health Data Hub est une plateforme qui vise d'une part à centraliser une grande partie des données de santé des personnes en France et d'autre part de permettre leur exploitation algorithmique à des fins de recherche tant publique que privée. Le CCNE et le Comité National Pilote d'Éthique du Numérique ont émis un avis « Plateformes de données de santé : enjeux d'éthique » [50].

Côté éducation, les États Généraux du Numérique pour l'éducation organisés en novembre 2020 par le Ministère de l'Éducation Nationale, de la Jeunesse et des Sports ont retenu un certain nombre de propositions [51], dont la proposition 37 qui préconise de « Créer l'« *Education Data Hub* », la plateforme de données d'éducation », pour faire avancer la recherche en éducation, éclairer les décisions, construire des services plus performants. La proposition 36, quant à elle, propose d'« Intégrer le programme européen GAIAX, cloud souverain pour l'hébergement des données de formation et d'éducation », afin de « construire une souveraineté européenne, dans un cadre d'interopérabilité et de sécurité commun ».

Il convient dès lors d'être vigilant sur le choix de l'hébergeur de la plateforme de données d'éducation afin de garantir la souveraineté européenne et d'assurer la protection effective des données.

Recommandation pour les Scientifiques - 12

— Comme les données de santé, les données pédagogiques sont fortement identifiantes et donc difficiles voire impossibles à anonymiser. On cherchera à suivre, lors de leur utilisation, des procédures

48. Avis n° 2020-1 du comité d'éthique pour les données d'éducation, *Enjeux d'éthique des usages des données numériques d'éducation dans le contexte de la pandémie*, recommandations 1 à 5 et 8 de la section « Respecter les libertés fondamentales des acteurs de l'éducation ».

49. https://www.monespacesante.fr/

50. CCNE : Comité Consultatif National d'Éthique pour les sciences de la vie et de la santé et CNPEN : Comité National Pilote d'Éthique du Numérique. *Plateformes de données de santé : enjeux d'éthique. Avis commun du CCNE et du CNPEN.* fr. Avis CCNE-143 et CNPEN-5. Fév. 2023. URL : https://www.ccne-ethique.fr/publications/avis-143-du-ccne-et-5-du-cnpen-plateformes-de-donnees-de-sante-enjeux-dethique.

51. https://www.education.gouv.fr/les-etats-generaux-du-numerique-pour-l-education-304117

similaires à celles qui sont utilisées pour les données de santé, dont en particulier la recherche systématique d'un consentement libre et éclairé.

Nous terminons cette section 5 par une recommandation transversale aux différents exemples de situations que nous avons examinés :

Recommandation pour les Pouvoirs Publics - 12

— Dès l'école primaire, puis dans le secondaire, concevoir un enseignement qui sensibilise aux enjeux de l'anonymisation, de la cybersécurité et de la maîtrise des outils permettant de préserver sa vie privée et la vie privée des autres, ainsi qu'aux enjeux de souveraineté.

6 En pratique, en tant que scientifique, comment dois-je procéder ?

Pour mener des recherches faisant intervenir des données personnelles, les scientifiques doivent savoir se questionner et effectuer des traitements que nous analysons maintenant. Il y a en effet un compromis à trouver entre l'utilité des données et leur anonymisation, en raison de leur appauvrissement inéluctable lorsqu'un processus d'anonymisation leur est appliqué.

Compte tenu de la sensibilité du sujet des données personnelles, de leur statut spécifique relativement à la recherche scientifique, à l'évolution de la régulation et des lois encadrant ces sujets, la première recommandation, générique concerne l'encadrement statutaire de la recherche :

Recommandation pour les Scientifiques - 13

— Dès la préparation d'un projet de recherche mettant potentiellement en œuvre des données personnelles, le scientifique responsable du projet doit prendre contact avec le Délégué à la Protection des Données (DPD) (en anglais « *Data Protection Officer* » (DPO)) et le comité opérationnel d'éthique de la structure hébergeant la recherche.

Le cas échéant, le DPD pourra être amené à mettre en contact le scientifique responsable du projet avec le responsable de la sécurité des systèmes d'information (RSSI).

6.1 Quand les techniques d'anonymisation ou de pseudonymisation présentent un risque considéré comme acceptable

On se référera avec intérêt au document « Recherche scientifique (hors santé) : enjeux et avantages de l'anonymisation et de la pseudonymisation » publié par la CNIL [1] en janvier 2022, qui présente un certain nombre d'exemples et un tableau récapitulatif des principes de pseudonisation et d'anonymisation que nous reproduisons Figure 6.1.

Un autre document de la CNIL [2] indique les bases légales d'un traitement de recherche (hors santé) permettant à un organisme de traiter des données à caractère personnel.

6.1.1 Etape 1 : avant de travailler sur les données

6.1.1.1 Déterminer la finalité de son étude et les usages des données qu'elle nécessite

La traçabilité des données est très importante. Il est crucial de pouvoir déterminer quelle est l'origine des données et quelle est leur destination. Il est à noter que l'origine de ces données et la façon dont elles ont été obtenues peuvent également constituer des enjeux d'éthique, au-delà du risque de violer la vie privée des individus lorsque des données à caractère personnel sont utilisées.

1. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

2. https://www.cnil.fr/fr/recherche-scientifique-hors-sante-quelle-base-legale-pour-un-traitement-de-recherche, janvier 2022

PROCESSUS	PSEUDONYMISATION	ANONYMISATION
Statut des données	Personnelles : restent indirectement identifiantes et donc soumises au RGPD et à la Loi Informatique et Liberté	Anonymes
Réutilisation des données	Sous conditions	Sans restriction
Utilité des données	Préservée car pas d'altération du niveau de détail des données	Plus ou moins altérée en fonction des objectifs poursuivis et des méthodes appliquées
Méthodes à mettre en œuvre	Compteur, générateur de nombres aléatoires, fonctions de hachage, chiffrement à clé secrète, etc.	Dépend des objectifs poursuivis : confidentialité différentielle, randomisation, k-anonymat, ł-diversité, k-proximité, etc.
Complexité de la mise en œuvre	Simple à moyenne	Dépend des objectifs poursuivis : simple dans certains cas comme l'agrégation ou le comptage et complexe dans d'autres

FIGURE 6.1 – Principes de l'anonymisation et de la pseudonymisation - Tableau CNIL

1. D'où viennent les données que je prends pour démarrer mon étude ? Est-ce que je réutilise des données préexistantes ou des données publiques ? Si je réutilise des données, quelle est leur provenance ? Comment ont-elles été collectées : avec ou sans consentement ? et pour quel usage ? S'il s'agit d'une nouvelle collecte de données, quel processus ai-je suivi ?
2. Comment vais-je utiliser ces données ?
3. Quelles sont les obligations légales pour ces données personnelles ?
4. Pourquoi est-il souhaitable que j'utilise des données anonymes ?
5. Que vais-je faire des résultats que je vais obtenir à partir de ces données : s'agit-t-il d'un usage *a priori* interne ou ces résultats sont-ils directement utiles à d'autres travaux ? Est-ce que je vise à publier dans des revues avec nécessité de produire les jeux de données sous-jacents à l'étude pour permettre la relecture par des pairs et la reproductibilité ?

6.1.1.2 Evaluer si l'anonymisation est réellement une nécessité

1. Ai-je vraiment besoin d'information au niveau des individus ?
2. Ai-je suffisamment d'individus concernés pour garder les attributs intéressants pour l'étude ?
3. Suis-je d'accord pour perdre de l'information pour les finalités de l'étude ?
4. La pseudonymisation pourrait-elle suffire ?
5. Quels sont les coûts financier, environnemental et humain de la gestion de ces données personnelles ?

Comme indiqué dans le document de la CNIL [3], il y a quatre exceptions qui pourraient permettre le traitement de données personnelles même si elles sont sensibles dans le cadre de finalités de recherche,

3. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/focus-certaines-categories-donnees-personnelles

en précisant les exceptions mentionnées par le RGPD [4] :

(1) la personne concernée a donné son consentement explicite pour un ou des objectifs spécifiques ;

(2) les données sensibles sont manifestement rendues publiques par la personne concernée ;

(3) le traitement est nécessaire pour des motifs d'intérêt public important ;

(4) l'utilisation des données est nécessaire à la recherche publique.

6.1.1.3 Si je ne veux ou ne peux pas anonymiser

S'il ne veut pas perdre d'information et ne veut pas anonymiser, le chercheur peut travailler sur des données non anonymisées sous certaines conditions. Nous détaillons des solutions techniques disponibles à ce jour dans la section 6.2. Le chercheur devra tout de même contacter impérativement son DPD (Délégué à la Protection des Données) (*DPO, Data Protection Officer*).

6.1.1.4 Collecter des données

Le chercheur doit ici traiter du dilemme de la minimisation évoqué plus haut (voir section 2.3.2.3) et veiller aux éléments suivants :

1. Ne collecter que les données strictement nécessaires aux finalités de la recherche (principe de pertinence et de minimisation des données). Il est à noter que cet objectif de minimisation peut créer une tension avec les recherches qui nécessitent des données nombreuses, comme par exemple en apprentissage automatique.
2. Intégrer des scénarios d'attaque dans l'analyse préalable à la collecte des données (*privacy by design* [5]).
3. Développer une logique d'accès sécurisé et contrôlé.

6.1.2 Etape 2 : traitement des données

Selon les objectifs de son étude, le chercheur peut alors recourir à un processus de pseudonymisation ou d'anonymisation. Il est important de noter que les processus de pseudonymisation et d'anonymisation sont des traitements de données à caractère personnel au sens de l'article 4 du RGPD. Ils doivent donc être conformes aux règles de la réglementation en matière de protection des données à caractère personnel, comme n'importe quel autre traitement de telles données.

6.1.2.1 Pseudonymisation

La pseudonymisation (définie page 15) est un traitement de données personnelles réalisé de manière à ce qu'on ne puisse plus attribuer les données à une personne physique sans information supplémentaire. Ces informations supplémentaires créées lors de la procédure utilisée pour pseudonymiser permettent la réidentification en cas de besoin. Il convient, bien évidemment, de sécuriser le stockage de ces informations supplémentaires.

4. Voir aussi le document `https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf` qui présente le régime juridique applicable aux traitements de données à caractère personnel poursuivant une finalité de recherche scientifique (hors santé).

5. `https://www.enisa.europa.eu/publications/big-data-protection`

Dans sa fiche de 2022 [6] la CNIL donne des indications sur différentes techniques de pseudonymisation basées principalement sur l'utilisation de clés de chiffrement des informations.

On gardera présent à l'esprit le fait que les attaques par corrélations et croisements avec d'autres données permettent de ré-identifier des données pseudonymisées comme nous l'avons détaillé plus haut, et que celles-ci restent donc soumises au respect du RGPD.

6.1.2.2 Anonymisation

Pour réaliser l'anonymisation (définie page 13) différentes méthodes peuvent être utilisées et sont décrites section 4. Chaque solution doit être choisie et adaptée selon le type des données à anonymiser, la façon dont les données produites seront utilisées, les modèles d'attaques possibles, le coût des méthodes etc. Dans tous les cas, le chercheur doit limiter les risques de ré-identification. Il est à noter que selon la sensibilité des données, des solutions d'anonymisation plus ou moins robustes seront nécessaires. Par ailleurs, le type de publication envisagée est aussi important à considérer : des données publiées sur le Web, en libre accès, nécessiteront des garanties plus fortes que des données partagées avec un partenaire industriel avec qui un contrat juridique pourra être éventuellement signé [7].

Rappelons quelques techniques de base pour anonymiser :

— Ajouter du bruit (on modifie légèrement les valeurs des attributs) ;
— Permuter les valeurs d'attributs au sein des données ;
— Généraliser les valeurs pour former des groupes d'individus (k-anonymat, l-diversité etc.).

Dans tous les cas, évaluer le risque de ré-identification.

Un outil OpenSource développé par l'Université Technologique de Munich, ARX [8], permet de réaliser des anonymisations selon de nombreux modèles à partir de données originales au format tabulaire. En outre il fournit des méthodes d'évaluation de l'utilité des données, ainsi que des éléments d'appréciation des risques de ré-identification.

Nous rappelons les trois critères présentés en section 4.4.1 qui permettent d'assurer que les données sont vraiment anonymes : non-individualisation, non-corrélation et non-inférence.

6.1.3 Etape 3 : après le traitement

— Réaliser une analyse d'impact sur la protection des données (AIPD) en application de l'article 35 du RGPD [9]. La CNIL met à disposition du public son logiciel open source PIA (*Privacy Impact Assessment*) [10] (version 3.0 en juin 2021 [11]) qui facilite la conduite et la formalisation d'analyses d'impact relatives à la protection des données.
— Documenter les techniques utilisées : le chercheur est responsable de la méthode choisie et appliquée, et le traitement qu'il a effectué doit pouvoir être réutilisé.
— Gérer le stockage et éventuellement le partage de ses données de façon sécurisée.
— En cas de publication des données, le chercheur doit s'assurer que la réutilisation de ses données pourra être faite de façon responsable [12].

6. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

7. Nguyen et Castelluccia, « Techniques d'anonymisation tabulaire : concepts et mise en oeuvre ».

8. https://arx.deidentifier.org/

9. https://www.cnil.fr/fr/RGPD-analyse-impact-protection-des-donnees-aipd

10. La CNIL indique que les termes AIPD (en anglais *Data Protection Impact Assessment, DPIA*), terme retenu dans le RGPD, d'une part, et PIA (*Privacy Impact Assessment*), terme plus commun utilisé dans d'autres régions du monde, d'autre part, sont synonymes.

11. https://www.cnil.fr/fr/analyse-dimpact-la-cnil-publie-la-version-3-0-de-son-logiciel-pia

12. Voir par exemple les critères du CIRAD pour la publication de ses données de recherche : https://coop-ist.cirad

— Avec le DPD, effectuer une veille des techniques de ré-identification et des sources de données susceptibles d'être croisées avec ses données de recherche à des fins de ré-identification.

6.2 Quand on ne peut pas anonymiser

6.2.1 Techniques homomorphes

Le chiffrement homomorphe permet de travailler directement sur les données chiffrées et évite d'avoir à transmettre ou à mettre à disposition des données non chiffrées. Le chiffrement homomorphe permet, par exemple pour l'opération d'addition, de rendre compatible cette opération avec une fonction de chiffrement au sens où la somme des chiffrés est exactement le chiffré de la somme des opérandes. Toute la confiance réside donc dans la qualité de l'algorithme de chiffrement. La difficulté dans ce contexte est la disponibilité de primitives de chiffrement homomorphes qui soient aussi universelles que possible au sens où elles soient homomorphes pour toutes les opérations nécessaires. Le chiffrement est dit totalement homomorphe lorsqu'il permet de prendre en compte toutes les fonctions calculables. Craig Gentry a montré en 2009 que de telles fonctions existent [13]. La complexité calculatoire des fonctions qu'il a mises en évidence est particulièrement élevée, ce qui ne permet pas actuellement une implantation réaliste de cette approche universelle. Des approches raisonnablement efficaces existent pour un nombre limité de fonctions telles que l'addition et la multiplication [14].

6.2.2 Données synthétiques

Lorsque le chercheur ne peut pas utiliser de données à caractère personnel car le processus d'anonymisation les rendrait inaptes à être utilisées dans sa recherche, par exemple, parce qu'elles seraient trop appauvries, il peut envisager d'utiliser des données synthétiques. Cette technique consiste à générer artificiellement des données qui ressemblent aux données réelles et à s'assurer que leur utilisation permet bien de valider les hypothèses souhaitées. Cette génération lui permet de pouvoir travailler avec des données tout de suite disponibles, souvent en aussi grande quantité que nécessaire et avec un spectre de valeurs aussi important qu'il le souhaite. Qui plus est, il connaît les différentes propriétés vérifiées par les données qu'il a générées, ce qui pourrait lui permettre d'éviter des biais dont pourraient souffrir des données recueillies auprès d'un échantillon non représentatif de personnes. Ces données, similaires aux données à caractère personnel, ne se rapportant pas à des individus existants, sont anonymes par définition et donc non soumises au RGPD, et pourraient remplacer des données réelles dans des conditions à analyser finement et à rendre explicites lors de la publication des résultats.

6.2.3 CASD - Centre d'accès sécurisé aux données

Actuellement plutôt méconnu des scientifiques, notamment du secteur public, le CASD (Centre d'Accès Sécurisé aux Données, cf. `casd.eu` , pour une description rapide) créé par arrêté interministériel du 29 décembre 2018, est un groupement d'intérêt public rassemblant l'État représenté par INSEE, le GENES, le CNRS, l'École polytechnique, HEC Paris et la Banque de France. Il a pour objet principal *« d'organiser et de mettre en œuvre des services d'accès sécurisé pour les données confidentielles à des fins non*

`.fr/gerer-des-donnees/rendre-publics-ses-jeux-de-donnees/3-rendre-publics-vos-jeux-de-donnees-est-une-decision-strategique`

13. Craig Gentry. « A fully homomorphic encryption scheme ». `crypto.stanford.edu/craig`. Thèse de doct. Stanford University, 2009.

14. Simon Fau et al. « Towards practical program execution over fully homomorphic encryption schemes ». In : *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2013)*. Compiègne, France, oct. 2013. url : `https://hal.archives-ouvertes.fr/hal-00917061`.

lucratives de recherche, d'étude, d'évaluation ou d'innovation, activités qualifiées de « services à la recherche », principalement publiques. Il a également pour mission de valoriser la technologie développée pour sécuriser l'accès aux données dans le secteur privé ». En pratique, le CASD est un tiers de confiance entre le producteur et l'utilisateur de données personnelles qui assure un stockage et un accès tout à la fois sécurisé et conforme à la réglementation européenne (RGPD).

Le CASD permet donc d'avoir accès à des données sensibles au sens du RGPD, en particulier à des fins de recherche. C'est une solution nécessitant de mettre en œuvre des protocoles spécifiques et contrôlés d'accès aux données et à leur manipulation.

6.2.4 Gérer des données sensibles

Gérer des données sensibles, dont les données personnelles, peut aussi être envisagé. Dans ce cas les scientifiques doivent mettre en œuvre les éléments suivants :

1. explicitation de la mise en œuvre du principe de proportionnalité ;
2. sécurisation appropriée, matérielle et logicielle des données ;
3. suivi de l'expérimentation par le DPD de l'établissement au titre duquel la recherche est effectuée et éventuellement par le Fonctionnaire Sécurité Défense (FSD) ;
4. archivage, sauvegarde et effacement des données en cours puis en fin de projet ;
5. attention toute particulière aux conditions de publication des résultats et à leur reproductibilité.

Les techniques citées dans les sections 6.1 et 6.2 ont besoin d'être améliorées. Nous recommandons de développer la recherche dans ces domaines.

Recommandation pour les institutions de recherche et d'enseignement - 4

— Développer des recherches sur :
(i) les techniques d'anonymisation ;
(ii) la minimisation des données personnelles : savoir déterminer l'ensemble approprié de données nécessaire à un travail scientifique ;
(iii) l'évaluation des risques de ré-identification sur les données anonymisées ;
(iv) la cybersécurité pour assurer des conditions de stockage et de partage sûres entre chercheurs ;
(v) des méthodes alternatives à l'anonymisation comme les techniques homomorphes ou la génération de données synthétiques.

6.3 Des documents

A- Général en France et à l'international

— Nous rappelons les trois documents de la CNIL édités en janvier 2022 et cités en section 6.1. Le premier fait le point sur les enjeux et avantages de l'anonymisation et de la pseudonymisation pour la recherche scientifique (hors santé) [15]. Le deuxième [16] rappelle les quatre exceptions qui pourraient permettre le traitement de données sensibles dans le cadre de finalités de re-

15. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/enjeux-avantages-anonymisation-pseudonymisation

16. https://www.cnil.fr/fr/recherche-scientifique-hors-sante/focus-certaines-categories-donnees-personnelles

cherche [17]. Enfin, le dernier [18] indique les bases légales d'un traitement de recherche permettant à un organisme de traiter des données à caractère personnel.

— Une note de l'agence nationale espagnole de la protection des données « Ten misunderstandings related to anonymisation » [19] (*Agencia espanola proteccion data* et EDPS) explique 10 malentendus relatifs à l'anonymisation.
— Le guide pratique sur l'anonymisation (60 pages) édité par la *Personal Data Protection Commission* (PDPC) de Singapour en mars 2022 illustre par des exemples diverses techniques d'anonymisation et propose un certain nombre d'outils d'anonymisation [20].
— Un article de 2023 du *National Institute of Standards and Technology* (NIST) fait le point sur les techniques d'anonymisation et de désidentification et donne une liste spécifique de quelques outils [21].
— Au Royaume Uni, un guide sur le UK-GDPR (*General Data Protection Regulation*) a été édité en janvier 2021 par le *Information Commissioner's Office* (ICO) [22].

B- Pour les données génomiques et de santé :

— Le rapport de l'agence européenne de médecine « Data anonymisation - a key enabler for clinical data sharing » [23] rassemble les discussions qui ont eu lieu sur le sujet lors d'un workshop qu'elle a organisé fin 2017.
— Le site de l'Alliance globale pour la génomique et la santé (*Global Alliance for Genomics and Health*, GA4GH [24].

C- En sciences sociales :

— L'InSHS (Institut des Sciences Humaines et Sociales, CNRS) a proposé en février 2021 une deuxième version de son guide pour la recherche « Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte » [25], illustrée d'exemples.

17. Un document plus ancien https://www.cnil.fr/sites/default/files/atoms/files/consultation_publique_-_presentation_du_regime_juridique_applicable_aux_traitements_a_des_fins_de_recherche.pdf présente les exceptions mentionnées par le RGPD.
18. https://www.cnil.fr/fr/recherche-scientifique-hors-sante-quelle-base-legale-pour-un-traitement-de-recherche
19. https://edps.europa.eu/system/files/2021-04/21-04-27_aepd-edps_anonymisation_en_5.pdf
20. https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Advisory-Guidelines/Guide-to-Basic-Anonymisation-31-March-2022.ashx, 60 pages
21. *De-Identifying Government Datasets : Techniques and Governance*, septembre 2023 : https://csrc.nist.gov/pubs/sp/800/188/final
22. https://ico.org.uk/media/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr-1-1.pdf
23. https://www.ema.europa.eu/en/documents/report/report-data-anonymisation-key-enabler-clinical-data-sharing_en.pdf
24. Framework for responsible sharing of genomic and health-related data, https://www.ga4gh.org/framework/
25. https://www.inshs.cnrs.fr/sites/institut_inshs/files/pdf/Guide_rgpd_2021.pdf

7 Synthèse des recommandations

7.1 Recommandation générale

1. Bien comprendre qu'on ne sait pas certifier aujourd'hui des procédés d'anonymisation. 4.4.2 p. 31
   L'expression «donnée anonymisée» ne devrait être utilisée que si le procédé d'anonymisation a été certifié.
   L'expression «donnée à faible risque de ré-identification» (DFRR) devrait être utilisée à l'issue d'une analyse d'impact relative à la protection des données (AIPD) concluant que le risque d'une possible ré-identification est maîtrisé et que les impacts potentiels sur la vie privée sont considérés comme faibles.

7.2 Recommandations pour les Scientifiques

1. Il ne serait pas éthique de demander, dans un protocole de recherche, à un sujet humain de consentir à renoncer à son anonymat. En général, il ne serait pas intègre de chercher à ré-identifier pour son propre usage scientifique des données personnelles. 2.3.4 p. 12
2. L'anonymisation n'est pas absolue et son efficacité dépend de facteurs évoluant dans le temps, comme la puissance des machines et les données disponibles. Il est nécessaire pour les scientifiques, les responsables de projets et d'unités de recherche d'adopter une approche dynamique, notamment en mettant en place une organisation interne appropriée afin de revoir régulièrement les techniques d'anonymisation et/ou de pseudonymisation utilisées au regard des avancées technologiques, de l'évolution des finalités poursuivies et des éventuelles catégories de données ajoutées dans un travail de recherche. 2.5.3.1 p. 17
3. Lorsque dans le cadre de ses travaux de recherche, un scientifique pense pouvoir s'appuyer sur l'Article 116 du décret n◦ 2019-536 du 29 mai 2019, il devra s'entourer de l'avis écrit du comité opérationnel d'éthique et de celui du Délégué à la Protection des Données (DPD) (en anglais le DPO) de son établissement, et s'assurer qu'il a obtenu toutes les autorisations requises par l'article 116 suscité. 3.2 p. 21
4. Dans leurs publications, les chercheurs doivent préciser les opérations effectuées (pseudonymisation, désidentification ou anonymisation), en se référant à la terminologie existante (RGPD et HIPPA). 4.2.2.1 p. 27
5. Mettre en place un groupe pluridisciplinaire avec des compétences techniques, éthiques et juridiques, pour élaborer un guide pour aider les chercheurs dans leur démarche d'anonymisation ou de désidentification de données textuelles. 4.2.2.1 p. 27
6. Partant du constat que la certification et l'homologation des procédés d'anonymisation sont aujourd'hui difficiles à mettre en place, il est recommandé aux scientifiques devant anonymiser des données, par exemple à des fins de publication, de faire valider leur procédé par le Délégué à la Protection des Données (DPD, en anglais DPO) de leur établissement. 4.4.2 p. 32
7. En matière de recherche, l'impératif éthique doit être adapté à chaque situation particulière, de 5.1.3 p. 35

manière à établir une relation de confiance entre les personnes dont on traite les données et celles qui y accèdent et les traitent. Il est essentiel que la personne dont on traite les données soit informée des modalités pour lesquelles l'autorité de contrôle assure sa fonction de tiers de confiance.

8. Informer les personnes dont on étudie les données génomiques de l'impact de cette étude à la fois pour elles mais aussi pour leur famille, leurs ancêtres et descendants, et plus généralement leur parentèle. 5.2.3 p. 39
9. Lorsqu'ils engagent des projets de recherche avec des données génomiques, les chefs de projets doivent expliciter les spécificités du consentement éclairé dans ce contexte. 5.2.3 p. 39
10. Le consentement doit être adossé à des garanties relatives à la qualité scientifique des études réalisées avec les données génomiques, à leur caractère respectueux des personnes et aux mesures de gouvernance des bio-banques, en lien avec les comités d'éthique opérationnels locaux. 5.2.3 p. 39
11. Les chercheurs doivent veiller à ce que les données images qu'ils utilisent pour leurs recherches, en particulier celles qui sont disponibles sur les réseaux sociaux, vérifient les critères légaux requis. Il leur faut aussi s'assurer des conditions légales et respectueuses des personnes lors de la publication des résultats de recherche et de la diffusion des données permettant leur reproductibilité. 5.4 p. 42
12. Comme les données de santé, les données pédagogiques sont fortement identifiantes et donc difficiles, voire impossibles à anonymiser. On cherchera à suivre, lors de leur utilisation, des procédures similaires à celles qui sont utilisées pour les données de santé, dont en particulier la recherche systématique d'un consentement libre et éclairé. 5.5 p. 44
13. Dès la préparation d'un projet de recherche mettant potentiellement en œuvre des données personnelles, le scientifique responsable du projet doit prendre contact avec le Délégué à la Protection des Données (DPD) (en anglais « *Data Protection Officer* » (DPO)) et le comité opérationnel d'éthique de la structure hébergeant la recherche. 6 p. 46

7.3 Recommandations pour les pouvoirs publics

1. Il importe de se donner les moyens scientifiques, techniques et de régulation pour maîtriser les risques de ré-identification à partir de bases de données dont les identifiants ont été supprimés. 5.1.3 p. 35
2. Les professionnels de santé doivent bénéficier, lors de leur formation initiale et tout au long de leur carrière, d'une formation adaptée aux technologies numériques, aux principes éthiques qui régissent le recueil et le traitement des données, aux moyens à mettre en œuvre pour les respecter, et aux risques et biais qui résultent de leur non-respect. 5.1.3 p. 35
3. Compte tenu du rythme particulièrement important des innovations scientifiques et technologiques et des évolutions qu'elles déterminent dans le recueil et l'exploitation des données relatives à la santé, il est nécessaire d'évaluer périodiquement la mise en œuvre effective des dispositifs juridiques, afin de vérifier le maintien dans le temps de l'efficacité du système de protection des données personnelles qu'ils instaurent. 5.1.3 p. 35
4. Former les citoyens et en particulier les décideurs pour qu'ils comprennent les implications, en termes de confidentialité, du télé-versement de leurs données génétiques dans une base de données de généalogie génétique, surtout si elle propose un service de génétique « direct aux consommateurs ». 5.2.3 p. 39
5. Définir une stratégie nationale et portée par l'Europe concernant le développement de produits numériques d'éducation respectant les valeurs fondamentales Européennes. 5.5 p. 43
6. Sensibiliser les différents acteurs (enseignants, élèves, familles, entreprises, acteurs académiques, 5.5 p. 43

politiques) aux enjeux de souveraineté numérique dans le domaine des données scolaires numériques.

7. Engager une réflexion sur l'opportunité d'introduire dans le RGPD le statut de données sensibles pour les données pédagogiques, à l'instar des données de santé, à des fins de protection de la vie privée des élèves et des personnels de l'éducation. 5.5 p. 44
8. Donner un statut juridique plus protecteur pour les données d'éducation au niveau français en instituant des codes de conduite sectoriels sur les données d'éducation. 5.5 p. 44
9. Attirer l'attention des différents acteurs (enseignants, élèves, familles, entreprises, acteurs académiques, politiques) sur le caractère « spécifique » des données d'éducation et continuer à responsabiliser ces acteurs dans leur utilisation de ces données. 5.5 p. 44
10. Intensifier la formation au droit à la protection des données, en particulier de celles qui sont liées aux usages pédagogiques numériques, pour les enseignants, les élèves et leurs familles, en les illustrant avec des cas pratiques. 5.5 p. 44
11. Offrir des garanties de sécurité et de souveraineté des outils que l'Etat recommande ou met à la disposition des acteurs de l'éducation, et évaluer le risque que fait courir leur perte éventuelle de confidentialité. 5.5 p. 44
12. Dès l'école primaire, puis dans le secondaire, concevoir un enseignement qui sensibilise aux enjeux de l'anonymisation, de la cybersécurité et de la maîtrise des outils permettant de préserver sa vie privée et la vie privée des autres, ainsi qu'aux enjeux de souveraineté. 5.5 p. 45

7.4 Recommandations pour le grand public

1. Être conscient que l'anonymisation des données personnelles ne peut pas être certifiée actuellement sans une perte significative d'utilité et qu'un risque de ré-identification existe. 4.4.2 p. 32
   Être conscient du caractère évolutif et relatif des techniques d'anonymisation.

7.5 Recommandations pour les institutions de recherche et d'enseignement

1. Etant donné le caractère relatif et évolutif de l'anonymisation, il est souhaitable de se donner la capacité à renforcer la qualité et la sécurité des moyens de gestion des données mis à disposition des chercheurs par des mesures telles que : 2.5.3.1 p. 17
   (i) Renforcer et soutenir des initiatives qui visent à promouvoir les logiciels libres en développant des clouds maîtrisés ;
   (ii) Favoriser des solutions basées sur des systèmes interopérables pour une recherche fédérée ;
   (iii) Contribuer à mettre en place des infrastructures de collaboration et de mutualisation des données personnelles à des fins de recherche aux niveaux national, européen et international, selon les domaines de recherche, qui intègrent la protection de la vie privée.
2. Expliciter pour les scientifiques la portée du Considérant 26 du RGPD relatif aux données anonymes et à la notion de personne physique identifiable. 3.1.1 p. 19
3. L'information et la formation des chercheurs sur les questions d'anonymisation, de pseudonymisation ou de désidentification des données textuelles doivent être améliorées. 4.2.2.1 p. 27
4. Développer les recherches entre droit et numérique pour éclairer les évolutions possibles de la publication des données de justice. 5.3 p. 40

5. Développer des recherches sur : 6.2.4 p. 51

   (i) les techniques d'anonymisation ;

   (ii) la minimisation des données personnelles : savoir déterminer l'ensemble approprié de données nécessaire à un travail scientifique ;

   (iii) l'évaluation des risques de ré-identification sur les données anonymisées ;

   (iv) la cybersécurité pour assurer des conditions de stockage et de partage sûres entre chercheurs ;

   (v) des méthodes alternatives à l'anonymisation comme les techniques homomorphes ou la génération de données synthétiques.

8 Remerciements

Nos remerciements vont d'abord aux personnes suivantes qui ont été auditionnées ou consultées dans le cadre de la CERNA en 2019, ainsi qu'aux orateurs de la journée sur l'anonymisation des données de recherche à caractère personnel, organisée par la CERNA en juillet 2019 [1] :

— Thomas Baudel (IBM)
— Claude Castelluccia (Inria Grenoble)
— Vincent Delaitre (Deepomatic)
— Amandine Jambert (CNIL)
— Maryline Laurent (Télécom SudParis, Institut Polytechnique de Paris)
— Benjamin Nguyen (INSA Centre-Val de Loire)
— Jean Ponce (ENS-PSL)
— Bastien Rance (HEGP, Université Paris-Cité)
— Vincent Rivollier (Université Savoie-Mont-Blanc)
— Josef Sivic (Inria Côte d'Azur)
— Alain Viari (Inria Grenoble)
— Pierre Zweigenbaum (LISN Orsay)

Comme nous l'avons précisé dans le préambule, nous avons initié ce travail au sein d'un groupe de travail de la CERNA en 2018. Nous en remercions vivement les membres. Nous remercions aussi Alice René pour ses commentaires judicieux sur la version quasi-finale de ce document. Nos remerciements vont également à Anaëlle Martin pour sa relecture et ses suggestions avisées.

Toute erreur ou inexactitude reste évidemment de notre pleine responsabilité.

1. Journée CERNA Anonymisation des données de recherche à caractère personnel, 3 juillet 2019 : `http://cerna-ethics-allistene.org/`

9 Bibliographie

Abowd, John M. « The U.S. Census Bureau Adopts Differential Privacy ». In : *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*. 2018, p. 2867. URL : https://doi.org/10.1145/3219819.3226070.

Abul, Osman, Francesco Bonchi et Mirco Nanni. « Never Walk Alone: Uncertainty for Anonymity in Moving Objects Databases ». In : *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. ICDE '08. USA : IEEE Computer Society, 2008, p. 376-385. ISBN : 9781424418367. URL : https://doi.org/10.1109/ICDE.2008.4497446.

Acs, Gergely et Claude Castelluccia. « A Case Study: Privacy Preserving Release of Spatio-temporal Density in Paris ». In : *KDD '14 Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. Août 2014.

– « I Have a DREAM! (DiffeRentially privatE smArt Metering) ». In : *Proceedings of the Information Hiding Conference*. Août 2011.

Acs, Gergely et al. *Differentially Private Mixture of Generative Neural Networks*. 2018. arXiv : 1709.04514 [cs.LG].

Ács, Gergely, Gergely Biczók et Claude Castelluccia. « Privacy-Preserving Release of Spatio-Temporal Density ». In : *Handbook of Mobile Data Privacy*. Sous la dir. d'Aris Gkoulalas-Divanis et Claudio Bettini. Springer, 2018, p. 307-335.

Ahmed, Tanbir, Md Momin Al Aziz et Noman Mohammed. « De-identification of electronic health record using neural network ». In : *Scientific Reports* 10.1 (oct. 2020). ISSN : 2045-2322. URL : https://doi.org/10.1038/s41598-020-75544-1.

Article 29 Data Protection Working Party. *Avis 05/2014 sur les Techniques d'anonymisation*. Avr. 2014.

Brickell, Justin et Vitaly Shmatikov. « The Cost of Privacy: Destruction of Data-mining Utility in Anonymized Data Publishing ». In : *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08. Las Vegas, Nevada, USA : ACM, 2008, p. 70-78. ISBN : 978-1-60558-193-4. URL : http://doi.acm.org/10.1145/1401890.1401904.

Cao, Jianneng et Panagiotis Karras. « Publishing Microdata with a Robust Privacy Guarantee ». In : *Proc. VLDB Endow.* 5.11 (juill. 2012), p. 1388-1399. ISSN : 2150-8097. URL : http://dx.doi.org/10.14778/2350229.2350255.

CCNE : Comité Consultatif National d'Éthique pour les sciences de la vie et de la santé et CNPEN : Comité National Pilote d'Éthique du Numérique. *Plateformes de données de santé : enjeux d'éthique. Avis commun du CCNE et du CNPEN*. fr. Avis CCNE-143 et CNPEN-5. Fév. 2023. URL : https://www.ccne-ethique.fr/publications/avis-143-du-ccne-et-5-du-cnpen-plateformes-de-donnees-de-sante-enjeux-dethique.

CERNA. *La souveraineté à l'ère du numérique : Rester maîtres de nos choix et de nos valeurs*. fr. Rapp. tech. Nov. 2018, p. 36. URL : http://cerna-ethics-allistene.org/digitalAssets/55/55708_AvisSouverainete-CERNA-2018.pdf.

Chen, Bee-Chung et al. « Privacy-Preserving Data Publishing ». In : *Foundations and Trends in Databases* 2.1-2 (2009), p. 1-167. URL : https://doi.org/10.1561/1900000008.

Chevrier, Raphaël et al. « Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review ». eng. In : *Journal of medical Internet research* 21.5 (mai 2019).

Publisher: JMIR Publications, e13484-e13484. ISSN : 1438-8871. DOI : 10.2196/13484. URL : https://pubmed.ncbi.nlm.nih.gov/31152528.

CNPEN : Comité National Pilote d'Éthique du Numérique. *Enjeux éthiques des technologies de reconnaissance faciale, posturale et comportementale.* fr. Avis CNPEN-8. Nov. 2023. URL : https://www.ccne-ethique.fr/sites/default/files/2024-02/CNPEN_AVIS8.pdf.

Commission Nationale de l'Informatique et des Libertes (CNiL). *Methodology For Privacy Risk Management*. http://www.cnil.fr/fileadmin/documents/en/CNIL-ManagingPrivacyRisks-Methodology.pdf. 2012.

Dinur, Irit et Kobbi Nissim. « Revealing information while preserving privacy ». In : *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 9-12, 2003, San Diego, CA, USA*. 2003, p. 202-210. URL : https://doi.org/10.1145/773153.773173.

Domingo-Ferrer, Josep et Jordi Soria-Comas. « From t-closeness to differential privacy and vice versa in data anonymization ». In : *Knowl.-Based Syst.* 74 (2015), p. 151-158. URL : https://doi.org/10.1016/j.knosys.2014.11.011.

Dwork, Cynthia. « Differential Privacy ». In : *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*. 2006, p. 1-12. URL : https://doi.org/10.1007/11787006%5C_1.

Dwork, Cynthia et Aaron Roth. « The Algorithmic Foundations of Differential Privacy ». In : *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), p. 211-407. URL : https://doi.org/10.1561/0400000042.

Edge, Michael D. et Graham Coop. « Attacks on genetic privacy via uploads to genealogical databases ». In : *bioRxiv* (2019). DOI : 10.1101/798272.

El Emam, Khaled. *Guide to the de-identification of personal health information*. eng. Taylor & Francis, 2013. ISBN : 978-1-4665-7908-8.

Emam, Khaled El et Fida Kamal Dankar. « Research Paper: Protecting Privacy Using k-Anonymity ». In : *JAMIA* 15.5 (2008), p. 627-637. URL : https://doi.org/10.1197/jamia.M2716.

Erlich, Yaniv et Arvind Narayanan. « Routes for breaching and protecting genetic privacy ». In : *Nature Reviews Genetics* 15 (2014), p. 409-421. DOI : 10.1038/nrg3723.

Erlich, Yaniv et al. « Identity inference of genomic data using long-range familial searches ». In : *Science* 362.6415 (2018), p. 690-694. DOI : 10.1126/science.aau4832.

Fanti, Giulia C., Vasyl Pihur et Úlfar Erlingsson. « Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries ». In : *PoPETs* 2016.3 (2016), p. 41-61. URL : https://doi.org/10.1515/popets-2016-0015.

Fau, Simon et al. « Towards practical program execution over fully homomorphic encryption schemes ». In : *2013 Eighth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC-2013)*. Compiègne, France, oct. 2013. URL : https://hal.archives-ouvertes.fr/hal-00917061.

Fiore, Marco et al. « Privacy in trajectory micro-data publishing: a survey ». In : *Transactions on Data Privacy* 13 (2020), p. 91-149. URL : https://hal.inria.fr/hal-02968279.

Francis, Paul, Sebastian Probst Eide et Reinhard Munz. « Diffix: High-Utility Database Anonymization ». In : *Privacy Technologies and Policy - 5th Annual Privacy Forum, APF 2017, Vienna, Austria, June 7-8, 2017, Revised Selected Papers*. 2017, p. 141-158. URL : https://doi.org/10.1007/978-3-319-67280-9%5C_8.

Fung, Benjamin C. M. et al. « Privacy-preserving Data Publishing: A Survey of Recent Developments ». In : *ACM Comput. Surv.* 42.4 (juin 2010), 14:1-14:53. ISSN : 0360-0300. URL : http://doi.acm.org/10.1145/1749603.1749605.

GADOTTI, Andrea et al. « When the Signal is in the Noise: Exploiting Diffix's Sticky Noise ». In : *28th USENIX Security Symposium, USENIX Security 2019, Santa Clara, CA, USA, August 14-16, 2019*. 2019, p. 1081-1098. URL : https://www.usenix.org/system/files/sec19fall_gadotti_prepub.pdf.

GADOTTI, Andrea et al. « When the Signal is in the Noise: Exploiting Diffix's Sticky Noise ». en. In : *Proceedings of the 28th USENIX Security Symposium* (août 2019), p. 19. URL : https://www.usenix.org/conference/usenixsecurity19/presentation/gadotti.

GENTRY, Craig. « A fully homomorphic encryption scheme ». crypto.stanford.edu/craig. Thèse de doct. Stanford University, 2009.

GROUIN, Cyril, Nicolas GRIFFON et Aurélie NÉVÉOL. « Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs? » In : *Proceedings of the Sixth Int. Workshop on Health Text Mining and Information Analysis*. Jan. 2015, p. 31-39. DOI : 10.18653/v1/W15-2604.

GYMREK, Melissa et al. « Identifying Personal Genomes by Surname Inference ». In : *Science* 339.6117 (2013), p. 321-324. DOI : 10.1126/science.1229566.

HUKKELÅS, Håkon, Rudolf MESTER et Frank LINDSETH. *DeepPrivacy: A Generative Adversarial Network for Face Anonymization*. 2019. arXiv : 1909.04538 [cs.CV].

INSEE. *Guide du secret statistique*. 2018. URL : https://www.insee.fr/fr/statistiques/fichier/1300624/guide-secret.pdf.

LAVRENOVS, Arturs et Karlis PODINS. « Privacy violations in Riga open data public transport system ». In : nov. 2016, p. 1-6. DOI : 10.1109/AIEEE.2016.7821808.

LECLERQ, Pierre. « L'anonymat : une situation souvent légitime ; rarement un droit ». In : *Droit et technique : études à la mémoire du professeur Xavier Linant de Bellefonds*. Litec, impr., 2007. ISBN : 978-2-7110-0641-0.

LI, Ninghui, Tiancheng LI et Suresh VENKATASUBRAMANIAN. « t-Closeness: Privacy Beyond k-Anonymity and l-Diversity ». In : *Proceedings of the 23rd International Conference on Data Engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007*. 2007, p. 106-115. URL : https://doi.org/10.1109/ICDE.2007.367856.

LIN, Zhen, Art B. OWEN et Russ B. ALTMAN. « Genomic Research and Human Subject Privacy ». In : *Science* 305.5681 (2004), p. 183-183. DOI : 10.1126/science.1095019.

« Loi nr 51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques ». In : *Journal Officiel de la République Française* 6 juin 1951 (1951), p. 6013. URL : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000888573&dateTexte=20190909.

« Loi nr 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés ». In : *Journal Officiel de la République Française* 6 janvier 1978 (1978). URL : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=20190909.

MACHANAVAJJHALA, Ashwin et al. « *L*-diversity: Privacy beyond *k*-anonymity ». In : *TKDD* 1.1 (2007), p. 3. URL : https://doi.org/10.1145/1217299.1217302.

MEDICINE, Institute of. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC : The National Academies Press, 2015. ISBN : 978-0-309-31629-3. DOI : 10.17226/18998. URL : https://www.nap.edu/catalog/18998/sharing-clinical-trial-data-maximizing-benefits-minimizing-risk.

MEYERSON, Adam et Ryan WILLIAMS. « On the Complexity of Optimal K-anonymity ». In : *Proceedings of the Twenty-third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '04. Paris, France : ACM, 2004, p. 223-228. ISBN : 158113858X. URL : http://doi.acm.org/10.1145/1055558.1055591.

MEYSTRE, Stephane et al. « Automatic de-identification of textual documents in the electronic health record: A review of recent research ». In : *BMC medical research methodology* 10 (août 2010), p. 70. DOI : 10.1186/1471-2288-10-70.

MIR, Darakhshan J. et al. « DP-WHERE: Differentially private modeling of human mobility ». In : *BigData Conference*. 2013, p. 580-588.

MONTJOYE, Yves-Alexandre de et al. « Unique in the Crowd: The privacy bounds of human mobility ». In : *Scientific Reports, Nature* (mars 2013). URL : https://www.nature.com/articles/srep01376.

NAUTSCH, Andreas et al. « Preserving privacy in speaker and speech characterisation ». In : *Computer Speech & Language* 58 (2019), p. 441-480. ISSN : 0885-2308. DOI : https://doi.org/10.1016/j.csl.2019.06.001. URL : https://www.sciencedirect.com/science/article/pii/S0885230818303875.

NERGIZ, Mehmet Ercan, Maurizio ATZORI et Chris CLIFTON. « Hiding the Presence of Individuals from Shared Databases ». In : *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. Beijing, China : ACM, 2007, p. 665-676. ISBN : 978-1-59593-686-8. URL : http://doi.acm.org/10.1145/1247480.1247554.

NEY, Peter, Luis CEZE et Tadayoshi KOHNO. « Genotype Extraction and False Relative Attacks: Security Risks to Third-Party Genetic Genealogy Services Beyond Identity Inference ». In : *Network and Distributed System Security Symposium*. Jan. 2020. DOI : 10.14722/ndss.2020.23049.

NGUYEN, Benjamin et Claude CASTELLUCCIA. « Techniques d'anonymisation tabulaire : concepts et mise en oeuvre ». In : *1024 : Bulletin de la Société Informatique de France* 15 (avr. 2020), p. 23-41. URL : https://hal.archives-ouvertes.fr/hal-02570847.

NISSENBAUM, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2010.

PRASSER, Fabian et Florian KOHLMAYER. « Putting Statistical Disclosure Control into Practice: The ARX Data Anonymization Tool ». In : *Medical Data Privacy Handbook*. 2015, p. 111-148. URL : https://doi.org/10.1007/978-3-319-23633-9%5C_6.

RAISARO, Jean Louis et al. « Protecting Privacy and Security of Genomic Data in I2b2 with Homomorphic Encryption and Differential Privacy ». In : *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 15.5 (sept. 2018), p. 1413-1426. ISSN : 1545-5963. DOI : 10.1109/TCBB.2018.2854782.

RAJGURE, Sumit et al. « Reconstructing Obfuscated Human Faces with Conditional Adversarial Network ». In : *Machine Learning and Information Processing*. Sous la dir. de Debabala SWAIN, Prasant Kumar PATTNAIK et Pradeep K. GUPTA. Singapore : Springer Singapore, 2020, p. 95-104. ISBN : 978-981-15-1884-3.

« RÈGLEMENT (UE) 2016/679 DU PARLEMENT EUROPEN ET DU CONSEIL du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données)(Texte présentant de l'intérêt pour l'EEE) ». In : *Journal Officiel de l'Union Européenne* 2016/679 (2016). URL : https://eur-lex.europa.eu/eli/reg/2016/679/oj.

ROCHER, Luc, Julien M. HENDRICKX et Yves-Alexandre de MONTJOYE. « Estimating the success of re-identifications in incomplete datasets using generative models ». en. In : *Nature Communications* 10.1 (déc. 2019). ISSN : 2041-1723. DOI : 10.1038/s41467-019-10933-3. URL : http://www.nature.com/articles/s41467-019-10933-3 (visité le 10/01/2022).

SCHWARZ, Christopher G. et al. « Identification of Anonymous MRI Research Participants with Face-Recognition Software ». In : *New England Journal of Medicine* 381.17 (2019). PMID: 31644852, p. 1684-1686. URL : https://doi.org/10.1056/NEJMc1908881.

SIRDEY, Renaud, Thanhhai NGUYEN et Nabil BOUZERNA. « Authentification par reconnaissance de visages en chiffrement homomorphe ». In : *IEEE CloudCom*. Paris, France, avr. 2016. URL : https://hal.archives-ouvertes.fr/hal-01759524.

Stubbs, Amber, Michele Filannino et zlem Uzuner. « De-Identification of Psychiatric Intake Records ». In : *J. of Biomedical Informatics* 75.S (nov. 2017), S4-S18. issn : 1532-0464.

Stubbs, Amber, Christopher Kotfila et Özlem Uzuner. « Automated Systems for the De-Identification of Longitudinal Clinical Narratives ». In : *J. of Biomedical Informatics* 58.S (déc. 2015), S11-S19. issn : 1532-0464. url : https://doi.org/10.1016/j.jbi.2015.06.007.

Su, Jiawei, Danilo Vasconcellos Vargas et Kouichi Sakurai. « One Pixel Attack for Fooling Deep Neural Networks ». In : *IEEE Transactions on Evolutionary Computation* 23.5 (oct. 2019), p. 828-841. issn : 1941-0026. url : http://dx.doi.org/10.1109/TEVC.2019.2890858.

Sweeney, L. « Weaving technology and policy together to maintain confidentiality. » eng. In : *The Journal of law, medicine & ethics : a journal of the American Society of Law, Medicine & Ethics* 25.2-3 (1997). Place: England, p. 98-110, 82. issn : 1073-1105. doi : 10.1111/j.1748-720x.1997.tb01885.x.

Sweeney, Latanya. « Achieving k-Anonymity Privacy Protection Using Generalization and Suppression ». In : *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.5 (2002), p. 571-588. url : https://doi.org/10.1142/S021848850200165X.

Team, Differential Privacy. « Learning with privacy at scale ». In : 1.8 (2017). url : https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf.

« Version consolidée duu traité sur l'union européenne et du traité sur le fonctionnement de l'union européenne ». In : *Journal Officiel de l'Union Européenne* 2016/C 202/01 (2016). url : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=uriserv:OJ.C_.2016.202.01.0001.01.FRA&toc=OJ:C:2016:202:TOC#C_2016202FR.01001301.

Warren, Samuel D. et Louis D. Brandeis. « The right to privacy ». In : *Harvard Law Review* 4.5 (1890), p. 193-220. url : http://faculty.uml.edu/sgallagher/Brandeisprivacy.htm.

White, Tonya, Elisabet Blok et Vince D. Calhoun. « Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed ». In : *Human Brain Mapping* 43.1 (2020), p. 278-291. doi : https://doi.org/10.1002/hbm.25120.

Xiao, Xiaokui et Yufei Tao. « M-invariance: Towards Privacy Preserving Re-publication of Dynamic Datasets ». In : *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*. SIGMOD '07. Beijing, China : ACM, 2007, p. 689-700. isbn : 978-1-59593-686-8. url : http://doi.acm.org/10.1145/1247480.1247556.

Yang, H. et J.M. Garibaldi. « Automatic detection of protected health information from clinic narratives ». In : *Journal of Biomedical Informatics* 58 (déc. 2015). © 2015 Elsevier Inc. Made available under a Creative Commons license. http://creativecommons.org/licenses/by-nc-nd/4.0/, S30-S38. url : https://eprints.whiterose.ac.uk/108935/.

Index alphabétique