RÉPUBLIQUE DU NIGER
Fraternité - Travail - Progrès
MINISTÈRE DU PLAN
INSTITUT NATIONAL DE LA STATISTIQUE
PLATEFORME NATIONALE D'INFORMATION POUR LA NUTRITION

NIGER

MANUEL

JUILLET 2019

MANUEL SUR L'ANONYMISATION DES DONNÉES

OUTIL DE RENFORCEMENT DES CAPACITÉS EN ANONYMISATION DES DONNÉES

SOFRECO

PNiN
Plateformes nationales d'information pour la nutrition

HC3N
LES NIGERIENS NOURRISSENT LES NIGERIENS

Institut National de la Statistique
NIGER

AVANT-PROPOS

L'initiative « Plateformes Nationales d'Information pour la Nutrition (PNIN) », portée par la Commission Européenne, vise à aider les pays à renforcer leurs systèmes d'information et leurs capacités d'analyse de données pour la nutrition, de manière à mieux étayer les décisions stratégiques auxquelles ils sont confrontés pour prévenir la malnutrition et ses conséquences. L'approche développée par l'initiative PNIN consiste à renforcer les capacités des pays bénéficiaires du programme en matière d'exploitation optimale des données et informations existantes en lien avec la nutrition, de manière à ce qu'ils puissent mettre en œuvre des politiques et programmes efficaces et définir des priorités dans l'allocation des ressources avec l'appui des délégations locales de la Commission Européenne.

Dans le cadre de l'objectif spécifique 2 de l'Assistance Technique (AT) de la PNIN, à savoir « créer les capacités, au sein des parties prenantes au Niger, de formuler des questions/demandes en termes d'analyse, d'analyser les données afin de répondre à celles-ci et de mesurer les progrès effectués vers l'atteinte des objectifs nationaux de réduction de la prévalence de sous-nutrition », plusieurs formations ont été effectuées, dont deux formations en anonymisation des données. Ces formations doivent permettre de renforcer la diffusion de micro-données.

L'accès aux micro-données anonymes est un gage d'éthique respondant au principe de confidentialité des données. Ainsi l'anonymisation permet de renforcer la mission de service publique permettant aux utilisateurs de télécharger les bases de données pour des exploitations secondaires et approfondies. Toutes les données collectées par l'Institut National de la Statistique ou par les Directions Statistiques des Ministères possèdent des informations à caractère personnel, ne serait-ce qu'un identifiant auquel peut être attaché des informations plus ou moins sensibles. L'anonymisation répond à l'obligation de protéger les données pour différentes raisons morales. De plus, la législation nationale s'est durcie, des conséquences judiciaires ou pénales ne peuvent plus être écartées si un utilisateur peut identifier des individus sur une base mise en transparence. Éviter la ré-identification d'une donnée est plus compliqué qu'il n'y parait. En 2010, dans son étude « Uniqueness of Simple Demographics in the U.S. Population, Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000) », Latanya Sweeney a montré que quelques informations (sexe, date de naissance et commune) suffisent à identifier 87 % de la population américaine. Ainsi, l'anonymisation est une opération indispensable et irréversible qui consiste à transformer des données personnelles de façon à ne plus permettre l'identification.

Le présent manuel rappelle les principes de l'anonymisation (théorie et pratique) et doit permettre à l'INS et aux différents Services producteurs d'informations statistiques de protéger et accroître la notion du secret statistique et de confidentialité des données. S'inscrivant dans la suite des formations en anonymisation des données réalisées par Cominique Blum, consultant international, il rappelle tout l'intérêt porté par l'INS dans ce domaine depuis la mise en place d'une équipe chargée de la mise en œuvre, du suivi de l'anonymisation des données (Décision N°00000111 MP/INS/DG/DRH/DARC du 30 septembre 2019.

Guillaume **POIREL**
Expert Technique International, Chef de mission de l'Assistance Technique de la Plateforme Nationale d'Information pour la nutrition

SIGNALÉTIQUE

OURS

Unité responsable : Plateforme Nationale d'Information pour la Nutrition

Auteur : Dominique BLUM, Expert Court-Terme, consultant en anonymisation des données, Assistant Technique PNIN (AT/PNIN)

Editing : Guillaume POIREL, Chef d'Equipe, Statisticien-Analyste, Assistant Technique PNIN (AT/PNIN)

Editeur de la publication : Assistance Technique de la Plateforme Nationale d'Information pour la Nutrition / Institut National de la Statistique

PRÉAMBULE

« L'anonymisation » des données est un équilibre entre deux contraintes :

- Ne pas dénaturer ni appauvrir la « qualité scientifique » de la base de données, car les analyses statistiques qu'on veut pouvoir réaliser doivent être représentatives de la réalité ;
- Rendre les données inutilisables pour « un attaquant supposé » qui tenterait de reconnaître une ou plusieurs personnes afin de disposer à leur sujet des informations confidentielles qu'elles ont confiées au responsable de la base de données.

Anonymiser les données consiste donc à retirer des données les informations qui permettent de reconnaître les individus, tout en conservant suffisamment d'informations qui caractérisent ces individus.

On peut alors distinguer deux étapes dans l'opération d'anonymisation :

- Supprimer le risque « d'identification par les identifiants individuels » ;
- Contrôler et réduire le risque « d'identification sans identifiants individuels », généralement appelé « risque de ré-identification » bien que ce terme soit souvent impropre.

Mais même en respectant ces deux étapes, il arrive qu'on ne puisse pas se garantir totalement contre le risque qu'un attaquant potentiel parvienne à reconnaître un ou plusieurs individus. Alors on peut prévoir une troisième étape, qui ne relève pas à proprement parler de l'anonymisation : il s'agit du brouillage des informations sensibles.

SOMMAIRE

Avant-propos iii

Préambule 1

liste des figures 2

1. Les quatre catégories de données d'une base de données individuelles 5

1.1 IDENTIFIANTS INDIVIDUELS 5
1.2 DONNÉES SENSIBLES 5
1.3 QUASI-IDENTIFIANTS INDIVIDUELS 6
1.4 AUTRES DONNÉES 6
1.5 CATÉGORISATION PRÉALABLE DES DONNÉES 6

2 Supprimer le risque d'identification par les identifiants individuels, autrement dit supprimer le risque d'identification immédiat 7

2.1 ANONYMISATION AU SENS STRICT 7
2.2 PSEUDONYMISATION 7
2.2.1 Méthode de pseudonymisation n°1 : la numérotation séquentielle 8
2.2.2 Méthode de pseudonymisation n°2 : la numérotation aléatoire 8
2.2.3 Méthode de pseudonymisation n°3 : le hachage cryptographique 9

3. La ré-identification au sein d'une base de données pseudonymisée 12

3.1 EN QUOI CONSISTE LA RÉ-IDENTIFICATION ? 13
3.2 PSEUDONYMISER ET EMPÊCHER LA RÉ-IDENTIFICATION : DEUX PROCESSUS TRÈS DIFFÉRENTS 13
3.3 SUIVRE UNE DÉMARCHE PRAGMATIQUE 14
3.4 ESTIMATION EMPIRIQUE DU RISQUE INITIAL DE RÉ-IDENTIFICATION 14
3.5 LE CAS DE L'ÉCHANTILLONNAGE 16

4. Évaluer le risque de ré-identification : les métriques 17

4.1 *K*-ANONYMAT 17
4.2 *L*-DIVERSITÉ 17
4.3 LE RISQUE PROBABILISTE 20

5 Comment limiter le risque de ré-identification 23

5.1 LES MÉTHODES NON PERTURBATRICES .. 23
5.1.1 Supprimer les QIDs inutiles 23
5.1.2 Réduire le nombre de modalités des QIDs 24
5.1.3 Supprimer certaines valeurs des QID 24
5.1.4 Réduire la précision des données sensibles 25
5.2 LES MÉTHODES PERTURBATRICES 25
5.3 LA CRÉATION DE JEUX DE DONNÉES VIRTUELLES 26

LISTE DES FIGURES

Figure 1 : Extrait d'une base de données comportant des identifiants individuels 7
Figure 2 : Extrait de la base de données anonymisée 7
Figure 3 : Principe général de la pseudonymisation 8
Figure 4 : Pseudonymisation par numérotation séquentielle 8
Figure 5 : Pseudonymisation par numérotation aléatoire 9
Figure 6 : Exemple d'un message d'entrée fourni à la fonction de hachage cryptographique ... 10
Figure 7 : Exemple d'une clef de hachage fournie par la fonction de hachage cryptographique 10
Figure 8 : Extrait d'une feuille de tableur utilisant la fonction ComputeHash() 10
Figure 9 : Implémentation de la fonction ComputeHash() en VBA 11
Figure 10 : Distribution des modalités du QID « groupe d'âge » 14
Figure 11 : Tri décroissant des modalités du QID « groupe d'âge » 15
Figure 12 : Part cumulée des modalités du QID « groupe d'âge » 15
Figure 13 : Trois QIDs (forme, taille, contour) à deux modalités chacun, et une donnée sensible (couleur de fond) 18
Figure 14 : La combinaison de QIDs « rond – petit – contour rouge » détermine un groupe de 11 individus 18
Figure 15 : Malgré un effectif suffisant, l'absence de diversité du groupe n°4 le rend vulnérable 19
Figure 16 : Malgré un effectif suffisant, la faible diversité du groupe n°4 le rend vulnérable ... 19

1 LES QUATRE CATÉGORIES DE DONNÉES D'UNE BASE DE DONNÉES INDIVIDUELLES

La distinction qui suit entre les quatre catégories de données est nécessaire à la fois pour des raisons de clarté terminologique – ce sont ces appellations que nous utiliserons par la suite – et pour des raisons fonctionnelles : elles ne jouent pas le même rôle et ne nous occupent pas de la même manière dans les processus d'anonymisation.

1.1 IDENTIFIANTS INDIVIDUELS

Il s'agit d'informations identifiant directement ou indirectement les individus, de manière précise et le plus souvent presque parfaitement déterministe :

- Le nom patronymique et les prénoms en sont les exemples les plus évidents, mais pas les meilleurs puisqu'il est assez fréquent d'observer des homonymies complètes, et qu'on emploie généralement la date de naissance de manière complémentaire pour lever l'ambiguïté ;
- Les numéros d'identification individuels, rattachés à des personnes physiques, détenus par divers organismes ou administrations constituent de meilleurs exemples car ils sont affectés de manière absolument bijective[1] : numéro de sécurité sociale, numéro de passeport, numéro de carte nationale d'identité, numéro de permis de conduire, numéro d'assuré, numéro de mutualiste, numéro d'abonné, etc. ;
- Comme autres identifiants individuels, on peut également citer l'adresse de courriel privée, l'adresse postale complète de résidence, le numéro de téléphone fixe, le numéro de téléphone mobile, etc.

Bien entendu, dans les cas autres que le nom et le prénom, il faut disposer d'annuaires ou de « listes de correspondance » pour identifier nommément les personnes : correspondance entre le numéro de sécurité sociale et le nom des personnes, par exemple.

1.2 DONNÉES SENSIBLES

Il s'agit des informations confidentielles recueillies, qui relèvent le plus souvent de la vie privée des personnes ou des foyers concernés par la base de données. Ces informations ne sont généralement connues que d'un nombre restreint de personnes ou de structures, et leur caractère confidentiel est le plus souvent protégé par le secret professionnel ou le secret médical que doivent respecter ces personnes et ces structures : le niveau de revenus, la source des revenus, les éléments du patrimoine, les résultats scolaires, le détail des dépenses, les habitudes alimentaires, les infractions commises, les condamnations passées, les addictions, les comportements à risque, les pratiques sexuelles, les maladies guéries ou en cours de traitement, les médicaments consommés, les actes chirurgicaux subis ou prévus, les préférences politiques, philosophiques ou religieuses, etc.

La plupart du temps, ce sont ces données sensibles qui constituent la finalité du recueil de données considéré, et ce sont elles qui peuvent intéresser un attaquant potentiel : étant parvenu à reconnaître une personne dans la base de données grâce à ses identifiants individuels ou à d'autres procédés que nous évoquerons plus loin, c'est le contenu des données sensibles qui sera son trophée.

1 bijectivité : un numéro désigne un individu et un seul, et un individu est désigné par un numéro et un seul.

1.3 QUASI-IDENTIFIANTS INDIVIDUELS

Plus fréquemment évoqué par l'abréviation QID que par son intitulé complet, le quasi-identifiant est une catégorie d'information qui à elle seule elle ne permet généralement pas d'identifier un individu, mais qui combinée à d'autres QIDs y contribue plus ou moins fortement.

Ainsi, le code postal de résidence identifie le plus souvent un groupe de quelques centaines à quelques milliers d'individus. De même pour la date de naissance. Quant au sexe il partitionne la population en deux sous-ensembles de millions de personnes. Pourtant, des chercheurs ont pu démontrer que la combinaison de ces trois informations (code postal, date de naissance et sexe) détermine aux USA des groupes dont 87 % ne sont constitués que d'une seule personne ; autrement dit, dans près de neuf cas sur dix, on retrouvera à coup sûr l'enregistrement individuel (et son contenu sensible) dans une base de données qui comporterait notamment le sexe, la date de naissance et le code postal de résidence des personnes enregistrées.

Une caractéristique supplémentaire des QIDs réside dans le fait qu'il s'agit souvent d'informations « de notoriété publique », autrement dit connues de tous, et particulièrement d'un attaquant. Lorsque cela n'est pas le cas, il arrive qu'on puisse les obtenir dans d'autres bases de données dont le contenu est moins sensible. C'est ainsi par exemple qu'a procédé Latanya Sweeney en 1997 pour extraire le dossier médical (diagnostics, actes chirurgicaux, médicaments) du gouverneur du Massachusetts de la base de données médicales individuelles que celui-ci avait diffusée en la prétendant totalement anonymisée : pour quelques dollars, cette chercheuse du MIT avait tout simplement acheté la liste publique des électeurs inscrits dans l'État du Massachusetts, et y avait obtenu le code postal exact du gouverneur, identifié par son nom ; munie de ce code postal et de la date de naissance du gouverneur, de sexe masculin, elle n'avait eu aucune difficulté à retrouver son dossier médical.

D'autres informations que la date de naissance, le code postal et le sexe peuvent se révéler être des QIDs. Cela dépend essentiellement de la nature de la base de données considérée et de celle des attaquants potentiels. Par exemple dans une base de données médicales hospitalières, la date d'hospitalisation, la date de sortie, la durée de séjour constituent des QIDs, sous réserve bien entendu que l'attaquant les connaisse : un collègue de travail, un voisin, un membre de la famille, etc.

1.4 AUTRES DONNÉES

Pour la forme, il faut évoquer ici les informations disponibles dans une base de données, autres que les identifiants, les données sensibles et les QIDs. Ce sont généralement des données dont l'intérêt est exclusivement technique ou logistique : code de l'enquêteur, intitulé de l'enquête, date d'encodage, date de saisie, etc.

Ces informations ne jouent aucun rôle dans l'anonymisation ni la confidentialité des données.

1.5 CATÉGORISATION PRÉALABLE DES DONNÉES

Dans tout ce qui suit, nous supposerons que le travail préalable de catégorisation des informations de la base de données a été réalisé, de sorte que chaque champ de la base est classé dans la catégorie des identifiants individuels, des données sensibles, des QIDs ou des données techniques.

2 SUPPRIMER LE RISQUE D'IDENTIFICATION PAR LES IDENTIFIANTS INDIVIDUELS, AUTREMENT DIT SUPPRIMER LE RISQUE D'IDENTIFICATION IMMÉDIAT

L'anonymisation supprime la possibilité d'identification immédiate. Elle consiste à faire disparaître les identifiants individuels directs ou indirects de la base de données.

Mais en pratique il est exceptionnel qu'on procède ainsi sans un « filet de sécurité » : la plupart du temps, on remplace ces informations par une « étiquette » dans un processus qu'on appelle la pseudonymisation, cette « étiquette » devenant donc de fait un pseudonyme.

2.1 ANONYMISATION AU SENS STRICT

L'anonymisation au sens strict consiste à anonymiser sans pseudonymiser. Le processus est très simple : on supprime purement et simplement de la base de données toutes les « colonnes » classées dans la catégorie des identifiants individuels : noms, prénoms, numéros d'identification de tout genre, adresses personnelles de courriel, etc. La conséquence (recherchée) est qu'il devient impossible de retrouver immédiatement l'identité d'un individu de la base anonymisée.

Les figures 1 et 2 ci-dessous illustrent ce processus.

Figure 1 : Extrait d'une base de données comportant des identifiants individuels

Prénom	NOM	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	…
Mahamadou	ISSOUFOU	1952	Dandadji	Illéla	Tahoua	marié polygame	5	…
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	…
Guillaume	POIREL	…	…	…	France	marié monogame	3	…
…	…	…	…	…	…	…	…	…

Figure 2 : Extrait de la base de données anonymisée

date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	…
1952	Dandadji	Illéla	Tahoua	marié polygame	5	…
07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	…
…	…	…	France	marié monogame	3	…
…	…	…	…	…	…	…

2.2 PSEUDONYMISATION

La pseudonymisation a pour objectif d'obtenir une base de données anonymisée, ou plus exactement pseudonymisée, sans pour autant perdre le lien vers les individus concernés.

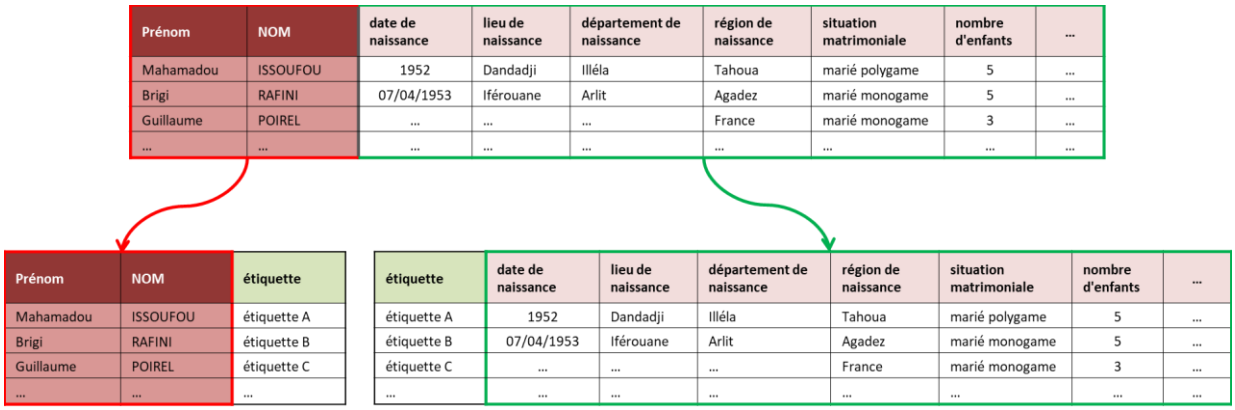
Cela n'a de sens que si personne ne détient simultanément la base anonymisée – qui contient les données sensibles – et l'ensemble de ces liens, qui constituent de fait une table de correspondance entre la base de données pseudonymisée et la liste des identifiants individuels.

Le principe général du processus de pseudonymisation revient à poser une étiquette individuelle, qu'on appelle pseudonyme, sur chacune des deux parties de chaque enregistrement de la base de données :

- Les identifiants individuels d'une part ;
- Les informations sensibles et les QIDs d'autre part.

Ce principe est illustré dans la figure 3 ci-dessous :

Figure 3 : Principe général de la pseudonymisation



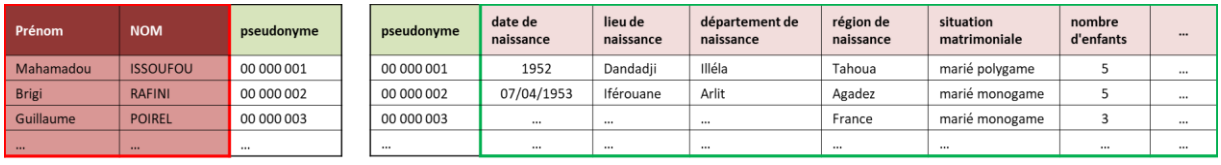
Si l'on exclut les méthodes manuelles, déconseillées et de toute manière inutilisables pour des bases de données conséquentes, on peut distinguer trois méthodes de pseudonymisation, dont le degré de complexité de mise en œuvre croissant va de pair avec les avantages qu'elles procurent : la numérotation séquentielle, la numérotation aléatoire, et le hachage cryptographique.

2.2.1 MÉTHODE DE PSEUDONYMISATION N°1 : LA NUMÉROTATION SÉQUENTIELLE

Très simple de mise en œuvre, cette méthode ne peut s'appliquer que si les données ne sont pas triées initialement selon un ordre naturel ou évident, qui anéantirait alors l'effort d'anonymisation en facilitant la ré-identification (ordre alphabétique des noms ; classement habituel des régions, des départements, des villes et villages ; ordre chronologique de la collecte ; etc.)

La figure 4 ci-dessous illustre cette méthode.

Figure 4 : Pseudonymisation par numérotation séquentielle



2.2.2 MÉTHODE DE PSEUDONYMISATION N°2 : LA NUMÉROTATION ALÉATOIRE

De mise en œuvre un peu plus complexe, cette méthode présente l'avantage de s'affranchir du problème engendré par les éventuels tris préalables, problème évoqué plus haut pour la méthode de numérotation séquentielle. Il faut bien entendu prévoir une plage de numérotation plus large que dans cette dernière.

La principale difficulté algorithmique de cette méthode est d'avoir à éviter de donner le même numéro à plusieurs enregistrements, sans devoir gérer une liste des numéros déjà attribués.

Une solution consiste à disposer d'un générateur de nombres aléatoires et à procéder en huit étapes successives, à l'issue desquelles le champ C contient le numéro aléatoire devant servir de pseudonyme :

1. Créer trois champs numériques supplémentaires A, B et C ;

2. Dans le champ A, numéroter la base de données de manière séquentielle (nécessaire pour l'étape 7) ;
3. Dans le champ B, attribuer à chaque enregistrement le nombre réel fourni par le générateur de nombres aléatoires auquel on impose une étendue à respecter (par exemple de 0 à 1), un nombre de décimales (par exemple 10) et à chaque itération une « graine » composée de divers champs de la base ;
4. Trier la base de données sur le critère du champ B, sans se soucier des *ex-æquo* ;
5. Dans l'ordre obtenu par ce tri, attribuer au champ C du premier enregistrement le numéro aléatoire n°1
6. En parcourant la base dans l'ordre de ce tri, attribuer au champ C de chaque enregistrement la valeur du champ C de l'enregistrement précédent augmentée d'un nombre entier fourni par le générateur de nombres aléatoires auquel on impose un minimum égal à 1, un maximum à respecter (par exemple 10) et un nombre de décimales égal à 0 ;
7. Trier la base de données sur le critère du champ A pour restaurer l'ordre initial ;
8. Supprimer les champs A et B.

Le résultat de cette méthode est illustré ci-dessous en figure 5.

Figure 5 : Pseudonymisation par numérotation aléatoire

Prénom	NOM	pseudonyme
Mahamadou	ISSOUFOU	870 145 032
Brigi	RAFINI	602 237 751
Guillaume	POIREL	537 501 223
...	...	...

pseudonyme	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
870 145 032	1952	Dandadji	Illéla	Tahoua	marié polygame	5	...
602 237 751	07/04/1953	Iférouane	Arlit	Agadez	marié monogame	5	...
537 501 223	...	...	...	France	marié monogame	3	...
...	...	...	...	...	...	...	...

2.2.3 MÉTHODE DE PSEUDONYMISATION N°3 : LE HACHAGE CRYPTOGRAPHIQUE

Le premier avantage de cette méthode est qu'elle permet de ne pas conserver la table de correspondance, ce qui renforce la protection. En effet, comme le hachage cryptographique garantit de toujours obtenir le même pseudonyme pour une même personne, il n'est pas nécessaire d'enregistrer le pseudonyme avec les données d'identification : en cas de besoin, on le recalculera « à la volée ».

Cette caractéristique explique le second avantage de la méthode : elle autorise l'appariement, parfois appelé « chaînage ». En d'autres termes, si deux bases de données non anonymisées comportent les mêmes champs pour identifier directement ou indirectement les individus, alors elles peuvent chacune être pseudonymisées par le même dispositif de hachage cryptographique, qui remplacera ces identifiants individuels par un pseudonyme identique.

Cela comprend donc :

- Le chaînage chronologique : collecte itérative concernant la même personne ;
- Le chaînage géographique : collecte concernant des personnes nomades ;
- L'appariement entre des bases comportant des informations complémentaires relatives aux mêmes individus ;

Le principe de la fonction de hachage cryptographique est le suivant :

- On fournit un « message en entrée » constitué d'une chaîne de caractères comportant les informations d'identification ;

- On applique à ce message la fonction de hachage cryptographique ;
- La fonction fournit en sortie une « clef de hachage » constituée d'une chaîne de texte sans signification composée de caractères alphanumériques :
  - ✓ dont la longueur fixe est « suffisamment grande » (par exemple 40 caractères)
  - ✓ parfaitement reproductible à partir du même message d'entrée
  - ✓ avec un taux de collision quasiment nul : deux messages distincts en entrée ne peuvent produire la même clef de hachage en sortie
  - ✓ avec un effet avalanche maximum : une modification infime du message en entrée produit une clef de hachage totalement différente en sortie
- Cette fonction n'est pas réversible : il n'existe pas de moyen pour reconstituer le message d'entrée à partir de la chaîne de sortie.

En pratique on utilise souvent en entrée un numéro d'identification individuel[2]. Les figures 6 et 7 ci-dessous fournissent une illustration du hachage cryptographique.

Figure 6 : Exemple d'un message d'entrée fourni à la fonction de hachage cryptographique

1 5 3 0 4 9 9 3 3 7 1 7 5 0 7 0 4 1 9 5 3 1

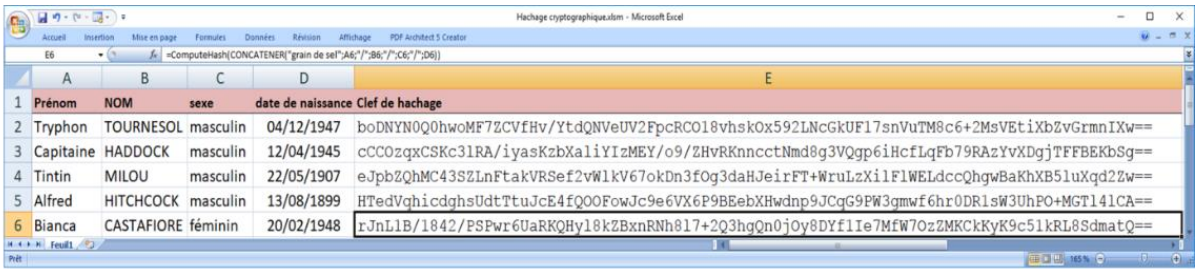
Figure 7 : Exemple d'une clef de hachage fournie par la fonction de hachage cryptographique

9pZ82kPH165bnUj3h075MmAx22yTM113FdF4n63967mAaS3TTz8582qeZwSQ45

La mise en œuvre du hachage cryptographique peut nécessiter des compétences informatiques. Toutefois il existe des fonctions de hachage dans le domaine public (MD5, SHA, …), certaines pouvant même être implémentées sous MS-Excel©.

Ainsi dans l'extrait de table MS-Excel© ci-dessous en figure 8, une fonction `ComputeHash()` a été implémentée dans le tableur et utilisée pour calculer le contenu de la colonne « E – Clef de hachage ».

Figure 8 : Extrait d'une feuille de tableur utilisant la fonction `ComputeHash()`



	A	B	C	D	E
1	Prénom	NOM	sexe	date de naissance	Clef de hachage
2	Tryphon	TOURNESOL	masculin	04/12/1947	boDNYN0Q0hwoMF7ZCVfHv/YtdQNVeUV2FpcRCO18vhskOx592LNcGkUF17snVuTM8c6+2MsVEtiXbZvGrmnIXw==
3	Capitaine	HADDOCK	masculin	12/04/1945	cCCOzqxCSKc3lRA/iyasKzbXaliYIzMEY/o9/ZHvRKnncctNmd8g3VQgp6iHcfLqFb79RAzYvXDgjTFFBEKbSg==
4	Tintin	MILOU	masculin	22/05/1907	eJpbZQhMC43SZLnFtakVRSef2vWlkV67okDn3fOg3daHJeirFT+WruLzXilFlWELdccQhgwBaKhXB5luXqd2Zw==
5	Alfred	HITCHCOCK	masculin	13/08/1899	HTedVqhicdghsUdtTtuJcE4fQOOFowJc9e6VX6P9BEebXHwdnp9JCqG9PW3gmwf6hr0DR1sW3UhPO+MGTl4lCA==
6	Bianca	CASTAFIORE	féminin	20/02/1948	rJnL1B/1842/PSPwr6UaRKQHyl8kZBxnRNh8l7+2Q3hgQn0jOy8DYf1Ie7MfW7OzZMKCkKyK9c51kRL8SdmatQ==

= ComputeHash(CONCATENER("grain de sel";A6;"/";B6;"/";C6;"/";D6))

Sous l'extrait, le zoom sur la formule utilisée dans la cellule E6 indique que le message en entrée

2 En France pour les données de santé par exemple, on utilise en entrée le « numéro de sécurité sociale », mais cela présente un inconvénient car souvent les enfants n'obtiennent ce numéro qu'après quelques années. La solution consiste alors à utiliser le numéro du parent assuré social et de le compléter par le sexe et la date de naissance du bénéficiaire des soins. Mais cette solution ne permet pas de distinguer les enfants jumeaux de même sexe.

est constitué de la concaténation du prénom, du nom, du sexe et de la date de naissance. La figure 9 quant à elle reproduit les quelques lignes de programme en VBA© qui ont été nécessaires pour implémenter cette fonction de hachage dans MS-Excel©.

Figure 9 : Implémentation de la fonction `ComputeHash()` en VBA

```
Function ToBase64String(rabyt)
  With CreateObject("MSXML2.DOMDocument")
    .LoadXML "<root />"
    .DocumentElement.DataType = "bin.base64"
    .DocumentElement.nodeTypedValue = rabyt
    ToBase64String = Replace(.DocumentElement.text, vbLf, "")
  End With
End Function

Function ComputeHash(fld)
    Dim text As Object
    Dim SHA512 As Object
    Set text = CreateObject("System.Text.UTF8Encoding")
    Set SHA512 = CreateObject("System.Security.Cryptography.SHA512Managed")
    ComputeHash = ToBase64String(SHA512.ComputeHash_2((text.GetBytes_4(fld))))
End Function
```

3 LA RÉ-IDENTIFICATION AU SEIN D'UNE BASE DE DONNÉES PSEUDONYMISÉE

La ré-identification est toujours une action illicite.

À proprement parler, le terme de ré-identification qu'on emploie couramment est meilleur que celui de désanonymisation. En effet la ré-identification consiste non pas à remettre un nom sur un enregistrement pseudonymisé, mais à retrouver l'enregistrement pseudonymisé d'une personne déterminée, ou d'un ensemble de personnes déterminées, afin d'obtenir des informations confidentielles à leur sujet.

Cependant ce terme est parfois inexact car dans certains cas, comme on le verra, un attaquant peut obtenir des informations confidentielles au sujet d'une personne ou d'un groupe de personnes sans pour autant avoir réussi à identifier avec certitude les enregistrements correspondants.

3.1 EN QUOI CONSISTE LA RÉ-IDENTIFICATION ?

Deux types d'attaque visent à ré-identifier une personne ou un groupe de personnes dans une base pseudonymisée :

- L'attaquant connaît déjà certains QIDs relatifs à une personne déterminée, et cherche à retrouver l'enregistrement ou les enregistrements qui lui correspondent au sein de la base de données ;
- L'attaquant a accès à d'autres bases de données dans lesquelles il dispose à la fois de combinaisons de QIDs et de l'identité des personnes concernées, et il les utilise unitairement ou en masse pour les apparier avec les QIDs de la base de données pseudonymisée.

Le premier type d'attaque serait plutôt le fait d'un proche de la personne ciblée, ou d'un attaquant visant ponctuellement une célébrité.

Le second type d'attaque est celui qu'a utilisé Latanya Sweeney en 1997 pour démontrer la faille de sécurité de la base de données médicales diffusée par le gouverneur du Massachusetts, comme évoqué plus haut. Ce type d'attaque serait plutôt le fait de *hackers* à la recherche d'exploits, d'organisations criminelles, de rançonneurs, etc.

3.2 PSEUDONYMISER ET EMPÊCHER LA RÉ-IDENTIFICATION : DEUX PROCESSUS TRÈS DIFFÉRENTS

Comme on l'a vu le processus technique nécessaire à la pseudonymisation d'une base de données est relativement simple à maîtriser, et consiste à faire disparaître les identifiants, informations inutiles pour le travail statistique proprement dit.

Mais la pseudonymisation n'est pas une opération suffisante pour supprimer le risque de ré-identification, puisque ce risque est lié au nombre et à la précision descriptive des QIDs, qui sont le plus souvent indispensables au travail statistique proprement dit, dont ils constituent généralement les axes d'analyse.

Empêcher la ré-identification consiste donc à réduire ou supprimer le risque lié à la rareté, voire à l'unicité des combinaisons de QIDs, sans supprimer ces QIDs. Le processus technique nécessaire pour y parvenir est plus complexe à maîtriser que celui de la pseudonymisation.

3.3 SUIVRE UNE DÉMARCHE PRAGMATIQUE

Il paraît assez évident qu'une base de quelques milliers de données individuelles, qui ne comporterait comme QID que le sexe des individus, ne permettrait pas à un attaquant de retrouver avec certitude dans la base les données concernant un individu déterminé[3].

Ainsi donc comporter des QIDs n'est pas suffisant pour engendre un risque réel de ré-identification d'une base de données : c'est le nombre de QIDs et le degré de finesse descriptive de ceux-ci qui déterminent en pratique l'importance du risque.

C'est pourquoi il est de bonne pratique de mesurer le risque de ré-identification avant toute chose. La démarche sera ensuite très simple :

- Tant que le risque dépasse un seuil acceptable, fixé à l'avance, on devra mettre en œuvre des opérations itératives de réduction du risque de ré-identification, jusqu'à le rendre acceptable ;
- Lorsque le risque est inférieur au seuil, il reste à régler les cas de « valeurs extrêmes des QIDs » (outliers), souvent très ré-identifiantes pour une poignée d'individus.

Comme cette démarche peut prendre du temps, on peut recourir à une méthode empirique d'estimation du risque initial de ré-identification, qui permet en quelques minutes de savoir si le risque est élevé et nécessitera un processus soutenu de réduction, ou non.

3.4 ESTIMATION EMPIRIQUE DU RISQUE INITIAL DE RÉ-IDENTIFICATION

La méthode que nous exposons ici est empirique mais suffisante, en général, pour évaluer la lourdeur du travail de réduction du risque de ré-identification. Elle fournit une valeur qui s'apprécie selon une fourchette elle-même empirique. Elle s'applique exclusivement à la base de données initiale, anonymisée mais dotée de tous ses QIDs avec leurs valeurs (modalités) d'origine non modifiées.

Pour illustrer la méthode, nous prendrons l'exemple d'une base de données de 10 000 individus.

La première étape consiste à établir la liste exhaustive des QIDs de la base de données. La nôtre en comporte trois : le groupe d'âge, le sexe et le type d'habitat.

Ensuite on trie les modalités de chaque QID par « part relative » décroissante, comme illustré dans les figures 10 et 11 pour le QID « groupe d'âge », et l'on calcule la part relative cumulée, de la modalité la plus fréquente à la moins fréquente, comme dans la figure 12. On relève alors le nombre de modalités nécessaires pour atteindre une part cumulée de 80% au moins, soit 4 dans notre exemple pour ce QID.

Figure 10 : Distribution des modalités du QID « groupe d'âge »

QID = Groupe d'âge	part relative
Nouveau-nés	17,2%
Enfants	25,7%
Adolescents	16,1%
Jeunes adultes	21,3%
Séniors	14,6%
Troisième âge	5,1%
total	100,0%

3 Sous réserve bien entendu que les deux sexes soient représentés équitablement dans cette base.

Figure 11 : Tri décroissant des modalités du QID « groupe d'âge »

QID = Groupe d'âge	part relative triée
Enfants	25,7%
Jeunes adultes	21,3%
Nouveau-nés	17,2%
Adolescents	16,1%
Séniors	14,6%
Troisième âge	5,1%
total	100,0%

Figure 12 : Part cumulée des modalités du QID « groupe d'âge »

QID = Groupe d'âge	part relative triée	part cumulée
Enfants	25,7%	25,7%
Jeunes adultes	21,3%	47,0%
Nouveau-nés	17,2%	64,2%
Adolescents	16,1%	80,3%
Séniors	14,6%	94,9%
Troisième âge	5,1%	100,0%
total	100,0%	

Les calculs étant effectués pour chaque QID, on effectue alors le produit des nombres obtenus pour tous les QIDs.

Supposons par exemple que :

- Pour le QID « sexe », les deux 2 modalités (« masculin » et « féminin ») sont nécessaires pour atteindre au moins 80 % de l'effectif total de la base ;
- Pour le QID « habitat », une seule des deux modalités (« rural », « citadin ») représente 80% de l'effectif total.

Alors le calcul du produit est le suivant : $produit = 4\,(pour\ «\ groupe\ d'âge\ ») \times 2\,(pour\ «\ sexe\ ») \times 1\,(pour\ «\ habitat\ »)$ soit $produit\ \ 8$

Ce produit est l'estimation empirique (et approximative) du nombre de combinaisons de QID distinctes observables dans la base de données.

Pour finir, on calcule l'estimateur empirique de l'effectif moyen d'individus pour une combinaison de QID déterminée, en divisant l'effectif total de la base par la valeur de ce produit, ce qui dans notre exemple donne : $estimateur\ empirique = 10\,000 / 8$ soit $estimateur\ empirique = 1\,250$.

On analyse alors ce résultat par rapport à trois seuils, déterminés empiriquement :

- Au-dessus de 30, le risque de ré-identification de la base de données est faible ;
- En-dessous de 10, le risque est fort ;
- Antre 10 et 30, le risque est moyen.

Comme indiqué plus haut, la conduite à tenir est alors la suivante :

- Au-dessus de 30, on se contente de régler le cas des *outliers*, s'il en existe : en effet même lorsque le risque global de ré-identification est faible, la distribution des QIDs peut présenter des valeurs extrêmes (donc rares) qui, combinées ensemble, désignent des individus très particuliers facilement reconnaissables (par exemple : homme, veuf, très jeune, avec 15 enfants)

- En-dessous de 30, on ne peut pas se contenter de cette estimation empirique : on doit alors procéder au calcul exact du nombre de combinaisons et de la distribution des effectifs des combinaisons et mettre en œuvre, si nécessaire, les procédés de réduction du risque exposés plus loin.

3.5 LE CAS DE L'ÉCHANTILLONNAGE

Tout ce qui vient d'être exposé s'applique parfaitement dans le cas d'enquêtes ou de bases de données exhaustives : le risque de ré-identifier une personne déterminée est bien égal au risque calculé puisque toute personne se trouve dans la base de données.

En revanche c'est différent dans le cas où la base de données est constituée par échantillonnage. Supposons en effet qu'un attaquant tente de trouver dans cet échantillon l'enregistrement correspondant à un individu déterminé : connaissant cet individu, l'attaquant connaît ses QIDs, et il va rechercher dans l'échantillon la combinaison de QIDs correspondante. Imaginons alors que cette combinaison ne se trouve qu'en un seul exemplaire dans cette base de données : rien ne prouve qu'il s'agit de l'individu recherché. En fait, si l'échantillonnage est aléatoire au taux de 1 pour 20, par exemple, on peut estimer qu'il y a une chance sur vingt pour que l'enregistrement soit celui de l'individu que recherche l'attaquant.

Ainsi, théoriquement, la meilleure dissuasion contre les attaques et les intrusions consiste à travailler avec des bases de données constituées par échantillonnage aléatoire.

Mais attention toutefois à plusieurs écueils :

- Le premier est l'excès de représentativité : dans un certain nombre de cas, les données sensibles ont une distribution très liée, statistiquement, à celles des QIDs ; en d'autres termes, cela signifie que les individus présentant le même profil de QIDs ont une probabilité très forte de présenter les mêmes valeurs de données sensibles, et dans ce cas l'attaquant peut se contenter de retrouver dans l'échantillon un individu similaire, par ses QIDs, à celui qu'il recherche, pour découvrir ses données sensibles avec une forte probabilité de succès ;
- Le deuxième écueil est l'excès d'information sur les modalités d'échantillonnage ; cela peut arriver notamment lorsque l'échantillonnage n'est pas absolument aléatoire : supposons par exemple qu'on décide de constituer un échantillon au 50ème, en n'intégrant dans la base de données que ceux dont le numéro de téléphone se termine par 25 ou 75. Il suffit que l'attaquant soit informé de cette méthode et il saura si la personne qu'il recherche – et dont il connaît le numéro de téléphone – fait partie de l'échantillon ou non. Alors le risque de ré-identification est soit nul (si le numéro de téléphone de la personne recherchée ne se termine ni par 25, ni par 75), soit égal au taux de ré-identification global (dans le cas contraire), l'ennui étant que l'attaquant sait précisément quels individus il est en mesure de ré-identifier ;
- Dernier écueil, proche du précédent, est l'excès de communication par les individus eux-mêmes : si l'enquête est valorisante pour les individus tirés au sort pour y participer, certains d'entre eux auront tendance à faire savoir qu'ils en font partie. Ce cas peut s'observer pour la constitution de cohortes d'individus auxquels on propose un bilan de santé gratuit, par exemple. Informé par l'individu lui-même (sur les réseaux sociaux, au cours d'une conversation en public, etc.) l'attaquant bénéficie alors pour cet individu du risque de ré-identification global.

Il faut donc garder à l'esprit que la protection contre le risque de ré-identification par l'échantillonnage aléatoire est relativement fragile, et peut parfois rassurer à tort les responsables de ces bases de données.

4 ÉVALUER LE RISQUE DE RÉ-IDENTIFICATION : LES MÉTRIQUES

Nous avons vu plus haut une méthode empirique d'estimation du risque de ré-identification dans une base de données. Cette méthode empirique a l'avantage d'être très rapide, mais elle débouche souvent sur la nécessité d'une évaluation plus précise, qui servira à piloter le processus de réduction du risque de ré-identification. Plusieurs métriques existent pour cette évaluation. Nous présentons ici les plus courantes, en deux catégories distinctes :

- Métriques globales : k-anonymat et l-diversité ;
- Métrique unitaire : le calcul probabiliste.

4.1 *K*-ANONYMAT

L'usage du k-anonymat pour estimer le risque global de ré-identification d'une base de données consiste à considérer que ce risque global est représenté par le groupe le plus vulnérable.

Le calcul du k-anonymat se réalise en trois étapes :

1. on distingue dans la base de données toutes les combinaisons effectives de QIDs ;
2. on compte l'effectif d'individus relevant de chacune de ces combinaisons ;
3. le k-anonymat est l'effectif le plus faible.

On interprète le résultat en fonction de seuils « habituels » :

- En-dessous de 5, c'est mauvais ;
- De 5 à 9, c'est médiocre ;
- Au-dessus de 10, c'est satisfaisant ;
- Au-dessus de 30, c'est très satisfaisant.

Utiliser le k-anonymat pour piloter le processus de réduction du risque de ré-identification incite donc à concentrer les efforts sur les groupes d'individus de plus faible effectif.

Le raisonnement sur lequel repose l'usage du k-anonymat peut se résumer de la manière suivante :

- Si le résultat est satisfaisant (k-anonymat > 10) c'est donc qu'aucun individu ne se trouve dans un groupe de moins de 10 personnes ;
- De sorte que même si un attaquant connaissait les QIDs relatifs à une personne déterminée, il ne pourrait pas la distinguer dans la base parmi les (au moins) 10 personnes présentant le même profil de QIDs (personnes « similaires ») ;
- Et ainsi il serait impossible pour un attaquant de retrouver les données sensibles relatives à un individu (ou à un ménage) en se fondant sur les QIDs relatifs à cet individu ou à ce ménage, dont il disposerait.

Cependant ce raisonnement est trop simplificateur, et c'est ce qui va justifier la nécessité d'un autre paramètre, la l-diversité.

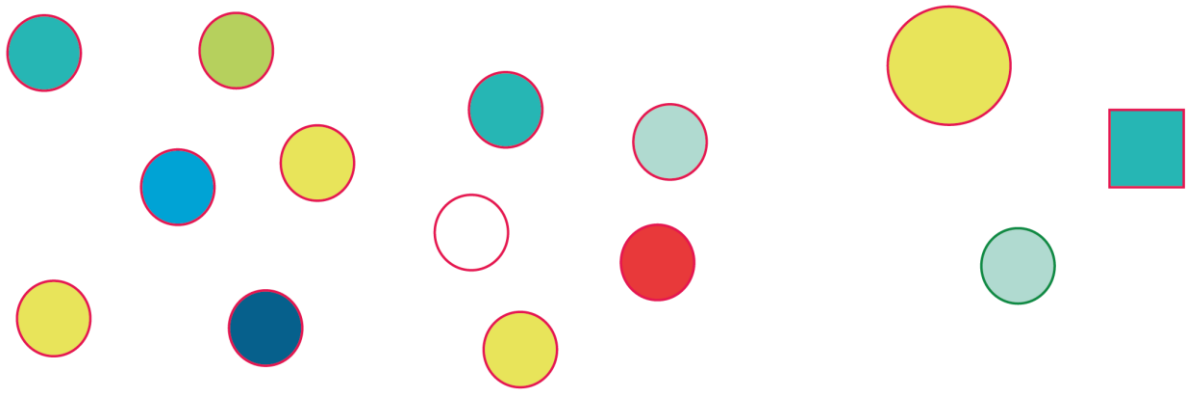
4.2 *L*-DIVERSITÉ

Les illustrations qui suivent vont permettre de comprendre le principe et l'intérêt de la l-diversité.

Nous allons les utiliser pour représenter graphiquement un ensemble d'individus dont les trois QIDs pourraient être le sexe (masculin ou féminin), la catégorie d'âge (adulte ou enfant) et

l'habitat (rural ou urbain), tandis que l'information sensible est le revenu. Dans nos illustrations, chaque individu est représenté par un objet dont la forme (rond ou carré), la taille (petit ou grand) et le contour (rouge ou vert) sont ces trois QIDs, tandis que la couleur de remplissage est l'information sensible.

Figure 13 : Trois QIDs (forme, taille, contour) à deux modalités chacun, et une donnée sensible (couleur de fond)

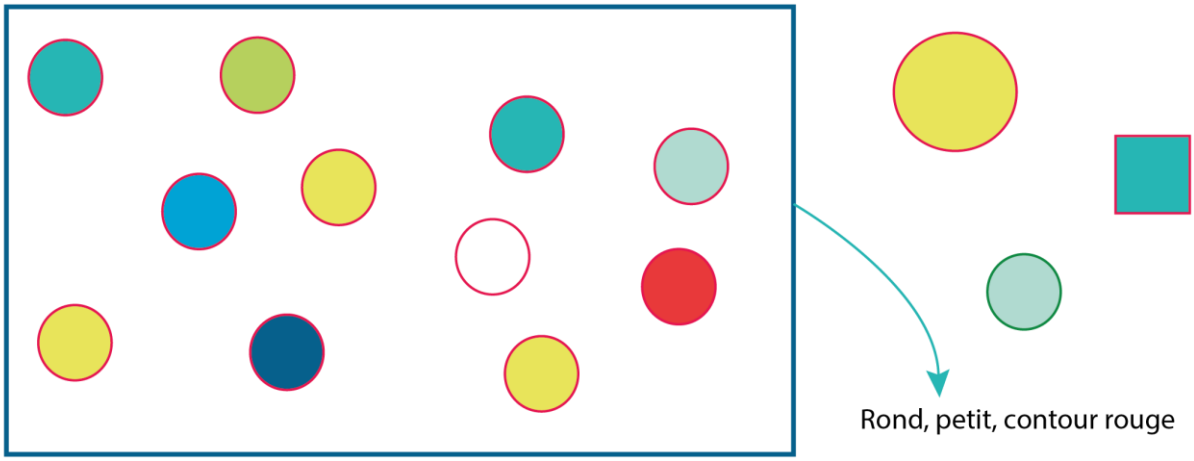


Notre base de données comporte donc quatorze individus. Si l'on énumère les trois QIDs de chacun d'entre eux, on constate que nous avons quatre groupes distincts :

- Groupe n°1 (1 individu) : rond, grand, contour rouge ;
- Groupe n°2 (1 individu) : carré, petit, contour rouge ;
- Groupe n°3 (1 individu) : rond, petit, contour vert ;
- Groupe n°4 (11 individus) : rond, petit, contour rouge.

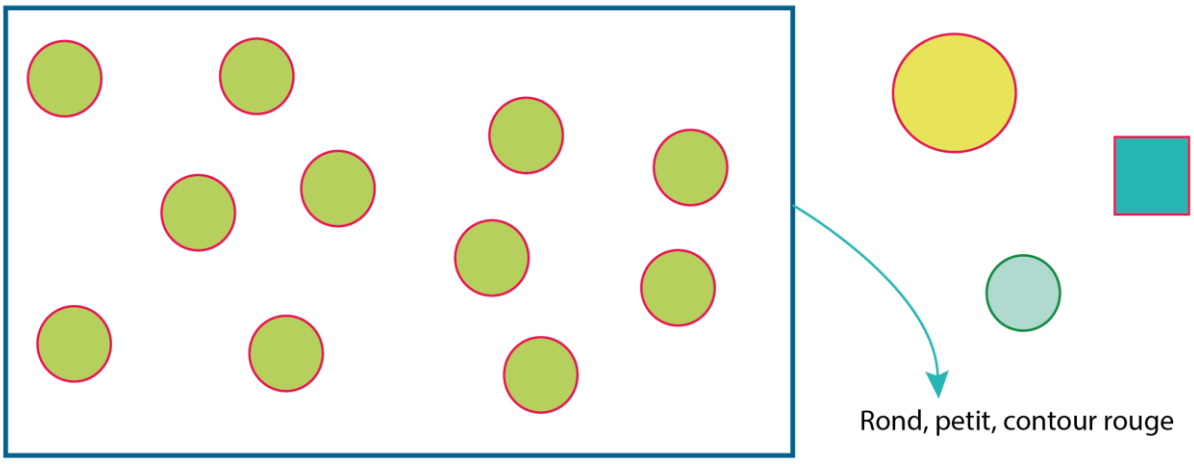
Nous allons nous intéresser plus spécialement au groupe n°4, qui comporte 11 individus. Cet effectif est supérieur à 10, de sorte que si un attaquant connaît un individu rond, petit et au contour rouge, il ne pourra pas le distinguer parmi les onze individus de ce groupe (figure n°14). Ainsi il ne pourra pas obtenir d'information sensible au sujet de cet individu, comme le montre la diversité des couleurs de remplissage représentées par ces onze individus.

Figure 14 : La combinaison de QIDs « rond – petit – contour rouge » détermine un groupe de 11 individus



Mais supposons à présent le cas illustré par la figure 15, dans lequel l'hétérogénéité des couleurs de remplissage des onze individus n'existe plus : tous ces individus sont de couleur vert pâle. Alors peu importe à l'attaquant de savoir lequel des onze individus précisément est celui qu'il recherche : tous les individus du groupe présentent la même information sensible, l'attaquant sait donc que l'individu qu'il recherche a le vert pâle comme donnée sensible.

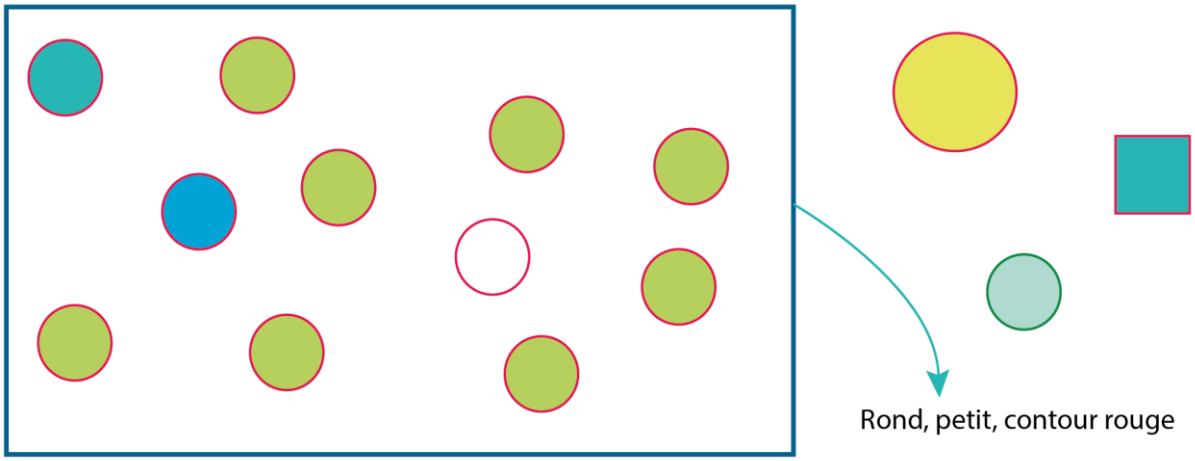
Figure 15 : Malgré un effectif suffisant, l'absence de diversité du groupe n°4 le rend vulnérable



Ainsi, malgré un effectif suffisant, l'absence de diversité du groupe n°4 l'a rendu vulnérable.

Et même sans aller jusqu'à ce cas limite, la figure 16 illustre une configuration dans laquelle le risque de dévoiler l'information sensible est élevé, puisque 8 individus sur les 11 que comporte le groupe présentent la même information sensible.

Figure 16 : Malgré un effectif suffisant, la faible diversité du groupe n°4 le rend vulnérable



Ces quelques illustrations graphiques ont mis en évidence la nécessité de compléter le k-anonymat pour estimer le risque global de ré-identification d'une base de données, par une mesure de la diversité des données sensibles dans les groupes constitués par les combinaisons de QIDs. C'est précisément l'objectif de la l-diversité.

Le calcul de la l-diversité se réalise en deux étapes complémentaires :

1. On compte le nombre de modalités distinctes de la donnée sensible au sein de chacun des groupes déterminés par les combinaisons de QIDs ;
2. La l-diversité est le nombre de modalités le plus faible.

On interprète le résultat en fonction de seuils « habituels » :

- En-dessous de 3, c'est mauvais ;
- De 3 à 5, c'est satisfaisant ;
- Au-dessus de 5, c'est très satisfaisant.

Dans les exemples illustrés plus haut la diversité du groupe n°4 est égale à :

- 10 dans la figure n°14 ;
- 1 dans a figure n°15 ;
- 4 dans la figure n°16.

De manière analogue au k-anonymat, l'usage de la l-diversité pour piloter le processus de réduction du risque de ré-identification incite donc à concentrer les efforts sur les groupes d'individus les moins hétérogènes, puisque la l-diversité représente l'hétérogénéité des données sensibles de la combinaison de QIDs la plus uniforme.

Il est rare, en pratique, que tous les groupes d'effectif suffisant (supérieur à 10) présentent une diversité égale ou supérieure à 3. C'est pourquoi, avant de lancer le processus de réduction du risque de ré-identification, il est utile de dénombrer les groupes « non conformes », l'effort à investir n'étant pas le même selon qu'on aura à régler le cas d'une dizaine de groupes, ou celui de plusieurs centaines de groupes.

Par ailleurs, il est fréquent que les bases de données comportent non pas une donnée sensible, mais plusieurs. On doit alors mesurer la diversité de chacune des données sensibles dans chacun des groupes, et il est convenu que la l-diversité de la base globale est celle de la variable sensible la moins diverse. Autrement dit, si la l-diversité de la variable A est égale à 5, et que celle de la variable sensible B est égale à 2, alors la l-diversité globale de la base est égale à 2.

4.3 LE RISQUE PROBABILISTE

- La méthode de calcul du risque que nous présentons ici peut paraître plus cartésienne et rigoureuse que l'utilisation du k-anonymat et de la l-diversité, néanmoins son application pratique est moins pertinente, notamment pour piloter le processus de réduction du risque de ré-identification. Elle est surtout utile pour relativiser le risque dans le cas d'un échantillonnage, sous réserve de ne rencontrer aucun des écueils évoqués plus haut.
- Le calcul est réalisé pour chaque combinaison de QIDs et ne fournit donc pas directement de mesure globale pour la base de données.
- Pour une combinaison donnée de QIDs, nous notons
  - ✓ X = nombre d'individus présentant cette combinaison dans la population ;
  - ✓ x = nombre d'individus présentant cette combinaison dans l'échantillon constitué par la base de données.
- Si l'on recherche l'enregistrement d'un individu présentant cette combinaison de QIDs, un simple raisonnement probabiliste indique la probabilité p de le trouver :
  - ✓ Cas où l'on sait que l'individu recherché se trouve dans la base de données : $p = \frac{1}{x}$
  - ✓ Cas où l'on ignore si l'individu recherché se trouve dans la base de données : $p = \frac{x}{X}$

Soulignons les inconvénients de cette méthode :

- Il faut calculer p pour chacune des combinaisons de QIDs observées dans la base de données ;
- En pratique, la plupart du temps on connaît x mais pas X, qu'il faut donc estimer ;
- La problématique de la diversité de la variable sensible n'est pas traitée.

5 COMMENT LIMITER LE RISQUE DE RÉ-IDENTIFICATION

Dès lors qu'une base de données pseudonymisée présente un risque de ré-identification, notre travail va consister à réduire ce risque, et si possible à le faire disparaître. Pour piloter ce processus de réduction du risque de ré-identification, la meilleure solution consiste à calculer les deux paramètres que nous avons vus précédemment (k-anonymat et l-diversité) de manière itérative, jusqu'à respecter les seuils fixés :

- k-anonymat supérieur ou égal à 10 ;
- l-diversité supérieure ou égale à 3

En d'autres termes, on poursuivra le processus tant qu'au moins un groupe comportera moins de 10 individus, et tant qu'au moins un groupe présentera une diversité inférieure à 3.

On dispose de trois grandes catégories de techniques de réduction du risque :

- Les techniques non perturbatrices ;
- Les techniques perturbatrices ;
- Les jeux de données virtuelles.

On peut parfaitement maîtriser les techniques non perturbatrices avec des outils statistiques et informatiques traditionnels. Les deux autres catégories en revanche font appel à des concepts mathématiques ou statistiques plus complexes, et contraignent à l'utilisation de logiciels spécialisés, qui sont des « boîtes noires » beaucoup plus délicates, pouvant alors fournir des résultats difficiles à comprendre, difficiles à expliquer, et potentiellement destructeurs. Nous recommandons de les utiliser avec parcimonie et uniquement si on en maîtrise tous les aspects.

Dans ce qui suit, nous nous intéresserons principalement aux techniques non perturbatrices, très efficaces en pratique courante et relativement simples à comprendre et à mettre en œuvre. Pour la forme, nous évoquerons succinctement les deux autres catégories de techniques.

5.1 LES MÉTHODES NON PERTURBATRICES

Dans cette catégorie de méthodes de réduction du risque de ré-identification, il y a schématiquement deux manières d'opérer, ces deux manières n'étant pas exclusives l'une de l'autre :

- **Jouer sur le k-anonymat** : réduire le risque de ré-identification en densifiant les groupes constitués par les combinaisons de QIDs, autrement dit en réduisant le nombre de combinaisons de QIDs. Dans ce cadre, on peut :
  - ✓ supprimer les QIDs inutiles ;
  - ✓ réduire le nombre de modalités de ces QIDs ;
  - ✓ supprimer certaines valeurs des QIDs.
- **Jouer sur la l-diversité** : réduire le risque de divulgation en réduisant la précision des données sensibles. Notons qu'il ne s'agit plus à proprement parler de techniques d'anonymisation.

5.1.1 SUPPRIMER LES QIDS INUTILES

Il s'agit d'une option à ne pas négliger. Les QIDs inutiles sont :

- D'une part des QIDs qui ont été recueillis mais dont on peut être sûr qu'on ne les utilisera dans aucune analyse : par exemple la couleur des yeux, la couleur des cheveux, la latéralisation (gaucher / droitier) dans une enquête sur les conduites alimentaires ;

- D'autre part des QIDs qui, combinés à d'autres, peuvent fournir une information utile tout en étant moins granulaire. Par exemple :
  - ✓ on peut supprimer la localisation du domicile et celle du commerce où se ravitaille la famille, en les remplaçant par le calcul de la distance domicile/commerce ;
  - ✓ on peut supprimer la date de naissance et la date de mariage, et les remplacer par l'âge au jour du mariage.

5.1.2 RÉDUIRE LE NOMBRE DE MODALITÉS DES QIDS

Cette technique est extrêmement efficace, car elle permet de réduire la granularité de certains QIDs, donc leur finesse descriptive. Bien entendu il ne faut pas la réduire de telle sorte qu'ils en deviennent inexploitables. Il s'agit donc de « flouter » les QIDs en regroupant plusieurs modalités.

Pour les QIDs quantitatifs (âge, poids, etc.) on utilise des plages de valeurs ; pour les QIDs catégoriels, on constitue des catégories plus larges.

Ce « floutage » (le terme consacré est « généralisation ») peut être exécuté soit sur l'ensemble de la base de données (on parle de généralisation globale) soit sur un sous-ensemble de celle-ci (généralisation locale ou différenciée). Dans le second cas cela produit au final un QID dont la précision est variable, ce qui peut rendre plus complexe les traitements statistiques ultérieurs.

Il est recommandé habituellement de commencer par généraliser les QIDs dont le nombre de modalités est le plus élevé, à moins que l'on dispose d'éléments particuliers commandant de procéder autrement. Ainsi prioritairement on généralisera

- Les localisations géographiques ;
- Les dates ;
- Les référentiels détaillés (professions, etc.) ;
- L'âge.

Le processus de généralisation est un processus itératif. Il est déconseillé de procéder à la généralisation de plusieurs QIDs dans la même itération.

Avant chaque itération, on calcule au moins les deux paramètres k-anonymat et l-diversité, puis on effectue les transformations de la variable que l'on veut généraliser, et l'on renouvelle enfin le calcul de k-anonymat et l-diversité : par comparaison avec les valeurs précédentes, on mesure l'impact de la généralisation à laquelle on vient de procéder.

Bien qu'efficace, la généralisation ne réalise pas des miracles : le plus souvent si l'on réduit le nombre de modalités par un facteur de 10, le nombre total de groupes de QIDs de la base ne sera divisé que par un facteur de 2, voire 3. En outre, réduire le nombre de modalités des QIDs agit très peu sur l'effectif de la combinaison de QIDs la plus rare (donc sur le k-anonymat), mais en revanche agit nettement sur les effectifs des combinaisons de QIDs les plus denses. C'est pourquoi on tentera lorsque c'est possible de mettre en œuvre une généralisation locale ou différenciée plutôt qu'une généralisation globale, malgré les difficultés qu'elle peut engendrer (modalités inhomogènes sur l'ensemble de la base).

5.1.3 SUPPRIMER CERTAINES VALEURS DES QIDS

En supprimant certaines valeurs de QIDs, qu'on traite alors comme des valeurs manquantes, on peut faire disparaître certaines combinaisons dont l'effectif trop faible va venir gonfler celui de groupes « fourre-tout ».

Il ne faut pas abuser de cette technique, et ne l'employer que pour des combinaisons bien déterminées (suppressions locales).

5.1.4 RÉDUIRE LA PRÉCISION DES DONNÉES SENSIBLES

Comme pour les QIDs, on peut « flouter » les données sensibles en les recodant par plages. Cependant il ne s'agit plus ici de limiter le risque de ré-identification, mais de délivrer une information sensible moins précise : on réduit donc le risque de divulgation en jouant sur la *l*-diversité.

Dans certains cas, on peut même supprimer purement et simplement certaines valeurs des données sensibles, soit pour toute la base (suppression globale), soit pour certaines combinaisons de QIDs seulement (suppression locale) ce qui conduit à les traiter ensuite comme des données manquantes.

5.2 LES MÉTHODES PERTURBATRICES

Contrairement à la généralisation qui transforme les données en respectant la plage de valeurs dans laquelle elles se trouvent initialement, les méthodes perturbatrices consistent à modifier les valeurs de certaines modalités, soit des QIDs, soit des variables sensibles, sans tenir compte des valeurs d'origine.

Par exemple :

- Alors qu'une généralisation de l'âge pourrait transformer la modalité « 3 ans » en « de 2 à 5 ans », une méthode perturbatrice pourrait transformer la modalité « 3 ans » en « 6 ans » ;
- Une méthode perturbatrice pourrait parfaitement transformer la modalité « masculin » en « féminin » pour le QID « sexe ».

Parmi les méthodes perturbatrices applicables aux QIDs, on peut retenir

- **Le bruitage** : ajout aléatoire d'une « petite » valeur positive ou négative à un QID quantitatif (l'âge, le poids, etc.) ;
- **La permutation** des valeurs d'un QID entre deux enregistrements distincts ;

Mais ces méthodes sont complexes à mettre en œuvre :

- Il faut préserver les moyennes, les variances, les corrélations etc. sans biaiser les résultats des études statistiques qui s'ensuivront ;
- Il ne faut pas créer des absurdités détectables, telles que :
  - ✓ une permutation « masculin / féminin » alors que la variable sensible est un diagnostic sexué (grossesse, tumeur de la prostate, …) ;
  - ✓ une permutation « adulte / enfant » alors que le statut matrimonial n'est pas lui-même permuté, l'enfant pouvant alors « devenir » marié ou veuf.

Ces deux méthodes perturbatrices sont également applicables aux données sensibles, et dans le cas d'une variable quantitative on peut également utiliser la méthode de la micro-agrégation : on attribue à tous les individus d'un groupe (déterminé par une combinaison de QIDs) la moyenne de cette variable pour ce groupe. Cette méthode, qui a pour conséquence paradoxale de réduire la diversité, permet de résoudre la cas de valeurs extrêmes ou rares, trop discriminantes individuellement (comme les revenus) sans modifier la position du groupe par rapport aux autres groupes.

5.3 LA CRÉATION DE JEUX DE DONNÉES VIRTUELLES

Nous citons cette méthode pour être exhaustif, mais elle est complexe à mettre en œuvre si l'on veut l'utiliser à des fins statistiques, le principe étant alors de créer des enregistrements cohérents, avec des moyennes, variances, corrélations, etc. conformes à la réalité.

En pratique cette méthode sera réservée plutôt à la fabrication de jeux de tests pour des chaînes de traitements informatiques à visée non statistique, nécessitant des volumes suffisants (tests de charge) et des cas suffisamment variés : dans ce cadre, il n'est pas nécessaire de respecter la cohérence ni la représentativité des données.

182, rue de la Sirba - BP 13416 Niamey - NIGER • Tél. : (227) 20 72 35 60 • Fax : (227) 20 72 21 74 • http://www.ins.ne •
http://www.pnin.org • E-mail : ins@ins.ne • E-mail : pnin@ins.ne