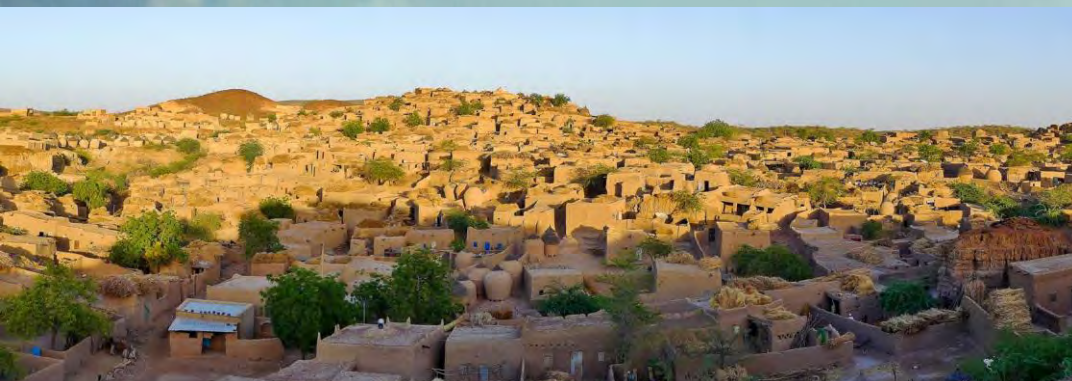


NIGER



UNION EUROPEENNE



RAPPORT DE FORMATION

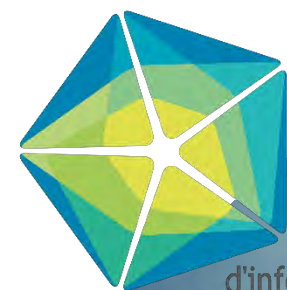
Seconde formation à l'anonymisation des données auprès de l'Institut National de la Statistique et des Directions Statistiques des Ministères Sectoriels

Plateforme Nationale d'Information pour la Nutrition au Niger

EuropeAid/139-061/DD/SER/NE

27-31 Mai 2019

A2055



PNiN

Plateformes nationales
d'information pour la nutrition



TABLE DES MATIERES

TABLE DES MATIERES	1
1. CONTEXTE ET OBJECTIFS DE LA MISSION	3
1.1. CONTEXTE DE L'ASSISTANCE TECHNIQUE A LA PNIN.....	3
1.2. CONTEXTE DE LA MISSION DE FORMATION	3
1.3. OBJECTIFS DE LA MISSION.....	4
2. DEROULEMENT DE LA MISSION	7
2.1. PREMIERE JOURNEE.....	7
2.2. DEUXIEME JOURNEE.....	8
2.3. TROISIEME JOURNEE.....	8
2.4. QUATRIEME JOURNEE	9
2.5. CINQUIEME JOURNEE	9
3. RECOMMANDATIONS	11
3.1. ORGANISATION DE LA FORMATION	11
3.1.1. <i>Organisation matérielle</i>	11
3.1.2. <i>Préparation des participants</i>	11
3.1.3. <i>Suggestion pour la diffusion de la formation</i>	12
3.2. EXPLOITATION PRATIQUE DES METHODES ET TECHNIQUES PRESENTEES	12
3.2.1. <i>Le logiciel miracle n'existe pas</i>	12
3.2.2. <i>L'évaluation empirique du risque de ré-identification est essentielle</i>	13
3.2.3. <i>Critères d'une anonymisation satisfaisante</i>	13
3.2.4. <i>Nombre de variables versus nombre d'individus</i>	13
4. ANNEXES	17
4.1. ANNEXE 1 : LISTES DE PARTICIPANTS – FORMATION EN ANONYMISATION DES DONNEES (27-31/05/2019)	19
4.2. ANNEXE 2 : CHRONOGRAMME DE LA FORMATION « ANONYMISATION DES DONNEES » 27-31/05/2019	21
4.3. ANNEXE 3 : SUPPORTS DE FORMATION	25

1. CONTEXTE ET OBJECTIFS DE LA MISSION

1.1. Contexte de l'Assistance technique à la PNIN

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 les plus concernés (Bangladesh, Niger, Burkina, Ethiopie, Laos, Kenya, Burundi, Zambie ...) 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.

L'initiative a pour but de produire de l'information liée à la nutrition, puis d'engendrer des besoins et demandes d'informations, de manière à alimenter le débat public et de reformuler des plans d'analyse pour les décideurs, les parties prenantes ou les partenaires de la nutrition.

L'Assistance Technique apporte principalement un appui technique et de renforcement de capacités liés aux résultats attendus du programme et qui doivent être déployés à différents niveaux institutionnels et décisionnels (INS, HC3N et Ministères Sectoriels).

1.2. Contexte de la mission de formation

Dans le cadre de l'atteinte de son objectif n°1 et plus particulièrement dans le souci de pouvoir collecter et traiter les données des Directions Statistiques des Ministères clés, l'assistance technique a souhaité mobiliser une Expertise Court Terme pour la formation en anonymisation des données.

Dans le cadre du programme PNIN, il est prévu de renforcer les capacités de l'INS et des Directions Statistiques de six ministères-clés (Santé, Education, Agriculture, Elevage, Hydraulique et Assainissement, Environnement). Lors de la phase d'élaboration du programme, puis de démarrage du projet, le besoin de renforcement des capacités pour l'anonymisation des données a été souligné, que ce soit au niveau de l'INS ou de certaines Directions Statistiques sectorielles.

L'accès aux micro-données anonymes est quasi-inexistant au niveau de l'INS. Pourtant, l'INS dispose d'une plateforme Anado (initiative de PARIS21, OCDE) qui devrait permettre aux utilisateurs et internautes de télécharger les bases de données de l'INS pour des exploitations secondaires et approfondies. Malheureusement, l'accès à ces bases de données ne peut se réaliser que si les bases de données sont anonymisées. L'absence d'anonymisation des données à ce jour ne permet pas de transmettre les bases de données en accès libre, seules les bases de données des Enquêtes Démographique et de Santé (EDS) semblent répondre aux attentes de l'INS dans le domaine de l'anonymisation des données.

Pour ce qui concerne la diffusion et la valorisation des données des secteurs, les mécanismes existants ont certaines faiblesses :

- Le transfert des données par les Directions Statistiques des secteurs et leurs validation auprès de l'INS sont inexistantes (seuls les annuaires sont transmis sous forme PDF) ;
- Les secteurs disposent de leur propre Comité de validation des données et n'ont pas l'habitude de transmettre leurs bases de données à l'INS (pourtant organe de coordination du Système Statistique National). Face à la réticence des Directions Statistiques sectorielles à transmettre les bases de données et à faire valider les informations statistiques par l'INS, s'ajoute la crainte de l'anonymisation des données et des risques d'identification des personnes, ménages enquêtés, structures.

Les Services producteurs d'informations statistiques responsables ont l'obligation de chercher à protéger les données. De plus, la législation nationale s'est durcie, des conséquences judiciaires ou pénales ne peuvent plus être écartées.

Dans ce cadre, il apparaît nécessaire de tenir une formation auprès des DS des ministères sectoriels, du HC3N et de l'INS pour approfondir le concept d'anonymisation des données et enseigner les différentes techniques pratiques dans le domaine.

1.3. Objectifs de la mission

L'objectif général est de parvenir à diffuser une information relative à la nutrition de qualité, vulgarisée, harmonieuse et accessible à tous. Plus spécifiquement, cette mission doit contribuer à :

- Protéger et accroître la notion du secret statistique et de l'anonymisation des données auprès des producteurs nationaux d'informations statistiques ;
- Mettre à disposition des utilisateurs les bases de données de l'INS et des Secteurs, que cela soit sur le Portail Anado de l'INS ou sur le futur portail de la PNIN.

Les techniques d'anonymisation doivent permettre ainsi :

- L'exploitation des données pour des traitements statistiques ;
- La création d'un jeu de tests réalistes pour les environnements hors-production ;
- la reconstruction du jeu de production sur un environnement pour étudier un incident.

2. DEROULEMENT DE LA MISSION

L'ensemble de la mission s'est déroulé en deux phases de formation de cinq journées. La seconde phase, qui fait l'objet de ce rapport, s'est déroulée du 27 Mai au 31 Mai 2019 à Niamey. Outre des représentants de l'INS, la formation a concerné des représentants du HC3N, des Ministères de l'Education, de l'Hydraulique et de l'Assainissement, de l'Environnement et de Développement Durable, de l'Agriculture et de l'Elevage et de l'Education Primaire, soit au total 14 personnes. La liste des participants est donnée en annexe 1.

Cependant la première journée (27 Mai) a été consacrée au bilan, après un an, de la formation dispensée lors de la première phase en Mai 2018, avec les participants de l'époque. Cette première phase s'était également déroulée à Niamey et avait concerné d'autres participants provenant de l'INS, du HC3N, du Ministère de l'Education Primaire (MEP) et du Ministère de la Santé Publique (MSP).

Une partie de la formation a été consacrée aux aspects théoriques, dont certains ont été ensuite mis en application lors d'exercices pratiques. Pour pouvoir en profiter pleinement, chaque participant disposait d'un ordinateur portable configuré avec son outil familier de traitement statistique (en pratique SPSS ou Stata). Dans la mesure du possible, les participants disposaient sur leur poste personnel d'extraits de bases de données réelles propres à leur activité, ou de structures de bases de données, sans données individuelles.

Toutes les demi-journées ont été ponctuées par des échanges avec les participants afin de répondre à leurs questions pratiques ou théoriques. Le chronogramme de formation est donné en annexe 2.

2.1. Première journée

La plupart des participants de la formation de 2018 ont pu s'organiser pour être présents.

Un tour de table est effectué, au cours duquel chacun expose le degré d'avancement des projets qu'il a en charge, l'utilisation qu'il a pu faire des concepts et techniques exposés l'an dernier et les difficultés qu'il a pu rencontrer.

Il apparaît que si les concepts sont bien assimilés, en revanche les applications sont rares. Le recul n'est sans doute pas suffisant, mais les participants évoquent aussi un besoin de travaux pratiques complémentaires.

La fin de la matinée et l'après-midi sont consacrés à un rappel méthodologique sommaire, à répondre à quelques questions techniques précises, à fournir des informations complémentaires utiles et à discuter de l'intérêt et de la place des outils logiciels « presse-bouton » qui commencent à exister et qui proposent de réaliser une anonymisation « clef en mains ».

2.2. Deuxième journée

Après une rapide présentation de ses attentes par chaque participant, le chef de la mission AT-PNIN rappelle le cadre et les objectifs de la formation.

La matinée est ensuite consacrée à l'approfondissement de la notion d'anonymisation : le sens habituel du mot « anonyme » est très insuffisant pour décrire le besoin d'anonymisation d'une base de données et il ne suffit pas de faire disparaître le nom et le prénom pour qu'une base de données devienne véritablement anonyme. On réalise alors que ce que l'on a pour habitude d'appeler « base de données anonymisée » est en réalité une « base de données pseudonymisée » et que le pseudonyme est en réalité un identifiant personnel indirect.

Les clefs et principes d'une bonne pseudonymisation sont présentées, afin de se familiariser avec un certain nombre de concepts complémentaires : base de données individuelles (microdata) versus base de données agrégées ; identification nominative directe ou indirecte ; pseudonymes versus quasi-identifiants ; variables d'identification versus variables sensibles.

La notion de risque de ré-identification est approfondie et décortiquée avec de nombreux exemples à l'appui. Les quasi-identifiants ou QID les plus courants sont passés en revue et un focus particulier est mis sur l'échantillonnage, une technique réputée réduire le risque de ré-identification.

La fin de l'après-midi est consacrée d'une part à la présentation d'une méthode empirique d'estimation du risque de ré-identification d'une base de données (cette méthode, certes grossière, permet en quelques minutes d'apprécier ce risque et de déterminer s'il est acceptable ou s'il convient de l'affiner et de le réduire) et d'autre part à l'exposé des différentes mesures de ce risque : k-anonymat, l-diversité et risque probabiliste.

2.3. Troisième journée

Au terme de la journée précédente, les participants s'étant approprié tous ces concepts ont réalisé pourquoi la formation ne s'achevait pas là. Contrairement à ce qu'ils auraient pu penser : non seulement la tâche qui consiste à pseudonymiser une base de données peut parfois présenter des difficultés, mais une fois pseudonymisée, il est encore possible dans la plupart des cas de ré-identifier les individus qui composent la base.

Des travaux pratiques sont effectués pendant toute la journée sur une base de données fictive préparée par le formateur, les participants n'ayant pas apporté leurs propres données. Ces travaux pratiques sont mis en œuvre sous MS-Excel® à

dessein : il s'agit de montrer que la plupart des techniques indispensables peuvent être réalisées avec un outillage standard de base :

- Distinction des différentes catégories de données (sous l'angle de l'anonymisation ;
- Suppression des identifiants ;
- Pseudonymisation par numérotation séquentielle ;
- Pseudonymisation par numérotation aléatoire ;
- Pseudonymisation par hachage cryptographique ;
- Mesure du k-anonymat ;
- Floutage de QIDs ;
- Impact du floutage sur le k-anonymat.

2.4. Quatrième journée

Toute la matinée est consacrée à la poursuite des travaux pratiques sur la base de tests :

- Mesure de la I-diversité ;
- Floutage de QIDs ;
- Impact du floutage sur la I-diversité ;
- Synthèse des travaux pratiques ;
- Formulation de requêtes SQL pour le calcul du k-anonymat et de la I-diversité.

Pendant l'après-midi, les techniques de réduction du risque de ré-identification sont passées en revue :

- Intervention sur les QIDs : floutage, suppression locale ;
- Intervention sur les données sensibles : suppression partielle d'information (échantillonnage, suppression d'enregistrements, floutage, traitement des outliers, suppression des valeurs discriminantes) et perturbation de l'information (création de bruit, permutation de données, micro-agrégation).

Une distinction est faite entre méthodes non-perturbatrices et méthodes perturbatrices. Les premières méthodes doivent être privilégiées car elles sont facilement compréhensibles par des tiers et de mise en œuvre relativement aisée, alors que les secondes nécessitent de recourir à des logiciels spécialisés qui fonctionnent en mode « boîte noire » et fournissent des résultats complexes à comprendre pour des non spécialistes.

2.5. Cinquième journée

Une partie de la journée est consacrée à récapituler la méthodologie complète :

- Données nominatives à supprimer ;
- Pseudonymisation rigoureuse ;
- Evaluation empirique du risque de ré-identification ;
- Stratégie : se satisfaire du résultat ou réduire le risque ;
- Réduire le risque en agissant sur les QIDs ;
- Réduire le risque résiduel en agissant sur les données sensibles ;
- Monitorer chaque étape avec les deux métriques : k-anonymat et I-diversité.

Quelques démonstrations sont réalisées pour répondre à des questions pratiques des participants : vérification des caractéristiques du hachage cryptographique, etc.

Sont ensuite abordées les questions posées par la diffusion d'une base de données personnelles, et à discuter de l'intérêt et de la place des outils logiciels « presse-bouton » qui commencent à exister et proposent de réaliser une anonymisation « clef en mains ».

La journée se conclut par une revue rapide de l'ensemble du programme parcouru depuis la première journée, donnant lieu à des échanges fournis et des recommandations pratiques.

3. RECOMMANDATIONS

En préambule, soulignons une caractéristique absolument délibérée de cette formation : il s'agit de faire en sorte que les participants puissent maîtriser les aspects théoriques essentiels de la démarche d'anonymisation et adapter leurs processus de production et de diffusion des bases de données personnelles sans avoir recours à des outils coûteux ni à des « boîtes noires » d'organismes tiers. Cela signifie que tout est fait pour que les participants parviennent à être autonome dans ce domaine en employant leurs outils familiers : SPSS, Stata, Excel, SQL, etc.

C'est sous cet éclairage qu'il convient de prendre en considération les recommandations qui suivent.

A l'issue de cette première session de formation, elles sont de deux types :

- Recommandations portant sur l'organisation de la formation
- Recommandations portant sur l'usage qui peut en être fait par les participants.

3.1. Organisation de la formation

3.1.1. Organisation matérielle

Les conditions matérielles sont parfaitement adaptées au déroulement de la formation et ne soulèvent aucune remarque particulière.

3.1.2. Préparation des participants

Un point a été nettement amélioré par rapport à la première session, la préparation des postes de travail individuels : cette fois-ci, nous n'avons pas perdu de temps à la préparation matérielle et logicielle, à l'installation des outils et à leur paramétrage initial.

En revanche, il subsiste toujours un problème pour la préparation de jeux de données de tests, car chaque organisme participant (chaque Ministère) est confronté à des problématiques spécifiques et seules des données issues de son activité habituelle aurait permis de faire émerger ces problématiques. Comme il ne peut pas être question d'apporter en formation des bases de données réelles (a priori non anonymes), il aurait été nécessaire que les participants préparent à

l'avance des jeux de données de test dans lesquels les informations auraient parfaitement pu être fictives, sous réserve que la structure des bases de données et les règles d'intégrité soient respectées.

En pratique, c'est sur un jeu de tests de 20 000 données fictives préparées par le formateur que les travaux pratiques ont pu être réalisés.

3.1.3. Suggestion pour la diffusion de la formation

Sous réserve d'une préparation plus efficace des participants (cf. point précédent), on peut envisager de réaliser la formation sur deux fois deux journées, au lieu de cinq. La troisième journée pourrait alors être consacrée à des exercices pratiques « hors site de formation » : exercices préparés durant les deux premiers jours par le formateur et adaptés à chaque participant, mais réalisés sans sa présence. La correction de ces exercices se déroulerait collectivement lors de la quatrième journée.

3.2. Exploitation pratique des méthodes et techniques présentées

3.2.1. Le logiciel miracle n'existe pas

Lors des échanges avec les participants, il est apparu que certains d'entre eux sont particulièrement en pointe sur la question de l'anonymisation des données et soumis à une forte pression pour produire des « bases de données anonymes ». Ils s'attendent donc à ce qu'on leur indique un logiciel miracle qui serait capable de prendre une base de données en entrée et de produire en sortie une base de données anonymes.

C'est un leurre fondé sur l'ambiguïté du terme « anonyme » : au sens propre, est anonyme une base de données dans laquelle les identifiants nominatifs n'apparaissent pas, ni directement ni indirectement. Qu'on les remplace par des pseudonymes numériques séquentiels, numériques aléatoires, ou résultant d'un hachage cryptographique, cela ne demande pas beaucoup de moyens informatiques. Mais en pratique supprimer les identifiants ne rend pas anonyme, puisqu'il existe des QID dont la combinaison peut permettre la ré-identification.

Ce leurre fait la fortune de quelques sociétés, mais cette idée qu'un « logiciel anonymisateur presse-bouton » peut résoudre en quelques instants les problématiques d'anonymat des données doit être combattue et abandonnée.

Il existe cependant des logiciels (ou des fonctions plug-in qu'on peut trouver sous R, par exemple) dotés de fonctionnalités spécialisées. Leur place est relativement étroite en application pratique, car ces outils complexes, s'ils sont mal maîtrisés, peuvent aboutir à rendre une base de données totalement inutilisable sans que cela ne soit aisément décelable. Ces outils sont en effet surtout focalisés sur les méthodes de masquage perturbatrices et se préoccupent peu, ou pas du tout, d'assister les utilisateurs pour réaliser une pseudonymisation parfaite (hachage cryptographique).

Pour autant il ne faut pas avoir de complexes : la problématique de l'anonymisation des données est très récente, elle se répand progressivement de manière universelle, mais aucun pays au monde ne prétend à ce jour avoir trouvé la solution miracle et tous les chercheurs spécialisés de ce domaine continuent à produire des méthodes, des techniques, des métriques, parfois des briques logicielles (API, webservices) mais pas de logiciel clef en main.

3.2.2. L'évaluation empirique du risque de ré-identification est essentielle

L'efficacité et la rapidité de la méthode empirique pour évaluer le risque de ré-identification en font un outil décisionnel très efficace, qui doit à ce titre être parfaitement maîtrisé.

3.2.3. Critères d'une anonymisation satisfaisante

Est satisfaisante une anonymisation qui atteint certains seuils pour le k -anonymat et la l -diversité, sans trop perdre de précision dans les données afin que la base de données conserve son « intérêt scientifique ».

Il est de bonne pratique de considérer comme satisfaisante une anonymisation qui aboutit à un **k -anonymat supérieur ou égal à 10 et une l -diversité supérieure à 3 dans plus de 98 % des combinaisons de QID**. Bien entendu, les combinaisons de QID présentant une diversité trop faible (moins de 2 % des combinaisons pour être satisfaisant) doivent soit être « noyées » dans des combinaisons plus larges, soit être exclues de la base de données finale.

Cependant il faut être vigilant pour la sélection de la variable sensible servant au calcul de la l -diversité, car un choix inapproprié peut se révéler artificiellement efficace : avec un k -anonymat égal à 10, si on n'observe aucune diversité inférieure à 3 dans les combinaisons initiales de QID, c'est sans doute que la variable sensible sélectionnée n'est pas la bonne.

3.2.4. Nombre de variables versus nombre d'individus

C'est le nombre de variables qui détermine le niveau de difficulté et finalement le temps qu'il faut consacrer à l'anonymisation satisfaisante d'une base de données :

- D'une part le nombre de variables ré-identifiantes (les QID), dont on connaît par ailleurs la difficulté liée à certaines d'entre elles, telles que la date de naissance exacte ou les coordonnées GPS : elles rendent l'exercice de l'anonymisation intégral particulièrement difficile, voire impossible ;
- D'autre part le nombre de variables sensibles : les bases de données étudiées pendant la formation de mai 2018 comportait de 250 à 300 variables sensibles distinctes, sur lesquelles il fallait appliquer les techniques de masquage pour éviter qu'une ré-identification aboutisse au dévoilement d'informations personnelles confidentielles (i.e sensibles).

Avec 8 à 10 QID « normalement discriminants » et 250 variables sensibles, il faut prévoir deux à trois mois de travail pour rendre une base de données

« **diffusable sans réserve** », c'est-à-dire avec un risque de ré-identification le plus faible possible et un risque quasiment nul de divulgation d'information personnelle sensible.

Le nombre d'individus dans la base de données influe sur les temps de traitement, mais c'est en général assez marginal par rapport au temps passé à réduire de manière itérative le risque de ré-identification ou de divulgation grâce aux différentes techniques (floutage, ...), à comparer les diverses tentatives, à retenir celle-ci plutôt que celle-là, etc.

L'échantillonnage est réputé faciliter la question de l'anonymisation : même si, connaissant un individu déterminé, on identifie une combinaison de QID unique qui lui correspond dans une base de données échantillonnée, on ne peut affirmer qu'il s'agit de cet individu car il n'a peut-être pas fait partie de l'échantillon. Cependant cela dépend d'une part du taux d'échantillonnage et d'autre part de la connaissance qu'on peut avoir des caractéristiques de la base de données complète : si l'on sait par exemple que 95 % des combinaisons de QID sont uniques dans la base complète, on peut être quasiment certain que la combinaison évoquée ci-dessus correspond effectivement à l'individu recherché. **Le cas de l'échantillonnage doit donc être considéré avec beaucoup d'attention, car il peut se révéler faussement rassurant.**

Finalement lorsque l'objectif est de parvenir à une diffusion sans risque de la base de données anonymisée, la conduite à tenir dont dépend le temps de traitement pour l'obtention de celle-ci, relève d'un des trois cas de figure suivants :

Premier cas de figure, défini par ces quatre conditions simultanées

- La base de données n'est pas exhaustive (c'est un échantillon de la population cible) ;
- ET le taux d'échantillonnage est suffisamment faible (au moins 1/100) ;
- ET personne ne connaît les caractéristiques de distribution des QID dans la population cible ;
- ET l'estimation empirique de l'effectif moyen de chaque combinaison de QID est égale ou supérieure à 10.

Alors la pseudonymisation est obligatoire. Le risque de ré-identification est quasiment nul et il faut simplement analyser la distribution des données sensibles afin de vérifier qu'aucune ne contient des outliers tellement exceptionnels qu'ils sont des « identifiants de fait » auquel cas il faut les masquer.

Avec 8 à 10 QID et 250 variables sensibles, ce traitement est obtenu en une semaine.

Deuxième cas de figure

- La base de données n'est pas exhaustive (c'est un échantillon de la population cible) ;
- ET le taux d'échantillonnage est moyennement faible (entre 1/100 et 1/20) ;
- ET personne ne connaît les caractéristiques de distribution des QID dans la population cible ;
- ET l'estimation empirique de l'effectif moyen de chaque combinaison de QID est égale ou supérieure à 10 ;

Alors la pseudonymisation est obligatoire, puis il faut traiter les données sensibles :

- Masquer les *outliers* ;
- Obtenir une diversité des données sensibles égale à 3 ou plus dans la totalité des combinaisons de QID, par les techniques de masquage des données sensibles.

Avec 8 à 10 QID et 250 variables sensibles, ce traitement demande quinze jours à deux mois de travail.

Autres cas de figure

- Echantillon avec un taux d'échantillonnage trop élevé (supérieur à 1/20) ;
- OU échantillon provenant d'une population dont les caractéristiques de distribution des QID sont publiques ;
- OU échantillon dont l'estimation empirique fournit une valeur inférieure à 10 pour l'effectif moyen de chaque combinaison de QID ;
- OU population cible exhaustive (ou quasi exhaustive).

Alors **la pseudonymisation est obligatoire** et il faut mettre en œuvre la totalité du processus de réduction du risque de ré-identification puis de traitement des données sensibles :

- Obtention d'un k-anonymat supérieur ou égal à 10 par les techniques de floutage (i.e généralisation) des QID ;
- Masquage des outliers ;
- Monitoring de la diversité des données sensibles les moins diverses dans le but d'atteindre une diversité au moins égale à 3 dans au moins 95 % des combinaisons de QID ;
- Réduction par les techniques de masquage du risque de divulgation des données sensibles.

Avec 8 à 10 QID et 250 variables sensibles, ce traitement nécessite un à trois mois de travail.

4. ANNEXES

4.1. Annexe 1 : Listes de participants – Formation en Anonymisation des données (27-31/05/2019)

N°	Prénom et Noms	Structure	Téléphone	Email
1	YAHOU Harissou	DS/MESUDD	96 48 22 67	yharissou2000@yahoo.fr
2	Moussa SaRkin GS	DCNCEE/INS	97 71 91 36	msgabass@ins.ne
3	Amadou Alhassane	DS/MSP	96 27 98 88	amadoualhassane@yahoo.fr
4	Amadou Mahamane	DS/MHA	90 10 51 53	amdjafarou@gmail.com
5	Ali Arzika	DER/INS	99 84 21 94	arzika@ins.ne
6	Harouna IDI	DER/INS	96 57 05 42	hidi@ins.ne
7	Mme Habibou Aminatou	DRFN/INS	96 55 78 97	haminatou@ins.ne
8	Nouhou Yahaya	DS/MEP	90 90 75 27	nouhou.yahaya@yahoo
9	Ibrahim Samaila Issa	DSEDS/INS	99 08 44 72	isamailai@ins.ne
10	Ali Ousmane	PNIN/INS	96 21 44 13	aousmane@ins.ne
11	Oumarou Zeinaba	DER/INS	91 04 84 84	ozeinaba@yahoo.fr
12	Poirel Guillaume	AT/PNIN	90 01 67 48	gpoirel@ins.ne
13	Maman Lawaly Boukari	PNIN/INS	96 32 33 02	lboukari@ins.ne
14	Gendah Neino	DS/MAG/EL	96 87 62 43	gondahn@yahoo.fr

4.2. Annexe 2 : Chronogramme de la formation « Anonymisation des données » 27-31/05/2019

Lundi 27 mai 2019 – INS du Niger	
9h00-9h15	Installation des participants (de la première session, mai 2018)
9h15-10h00	Tour de table : <ul style="list-style-type: none"> • applications concrètes depuis mai 2018 • quelles satisfactions ? • quelles difficultés ?
10h00 - 10h10	Pause « photo collective »
10h10-12h30	Rappel sommaire de concepts et de méthodologie (Dominique Blum) Questions des participants concernant les concepts et la méthodologie Réponses, compléments et approfondissement de certains points (Dominique Blum) Exemples pratiques
12h30-14h00	Déjeuner / Pause
14h00-17h00	Quels outils pour en faire quoi : tableau sommaire (Dominique Blum) Questions des participants concernant les outils employés et les outils manquants Réponses, compléments : la juste place des outils « prêts à porter » (Dominique Blum) Questions ouvertes et discussion
17h00	Échange de coordonnées


Mardi 28 mai 2019 – INS du Niger	
9h00-9h15	Installation des participants
9h15-10h00	Mot de bienvenue de Guillaume Poirel (Chef de mission AT PNIN) et de Dominique Blum (Expert en anonymisation des données) Présentation et définitions : anonymisation, pseudonymisation (Dominique Blum) Les catégories de données d'une base de microdata : identification nominative directe ou indirecte – Quasi-identifiants – Données sensibles (Dominique Blum) Questions des participants
10h00 - 10h10	Pause « photo collective »
10h10-12h30	Le concept de ré-identification d'une base pseudonymisée (Dominique Blum) Questions des participants Ré-identification par les QID (Dominique Blum) Questions des participants
12h30-14h00	Déjeuner / Pause
14h00-17h00	Estimation empirique du risque de ré-identification (Dominique Blum) Les différentes mesures de ce risque, et leur intérêt respectif (Dominique Blum) <ul style="list-style-type: none"> • méthode du k-anonymat et de la l-diversité • autres méthodes Questions des participants

Mercredi 29 mai 2019 – INS du Niger	
9h00-9h15	Installation des participants
9h15-10h50	Travaux pratiques sur bases de données de test : mesure empirique du risque de ré-identification (Dominique Blum) Questions des participants
10h50 -11h00	Pause
11h00-12h30	Travaux pratiques sur bases de données de test : mesure du k-anonymat avec SPSS, Excel ou SQL (Dominique Blum) Questions des participants
12h30-14h00	Déjeuner / Pause
14h00-17h00	Travaux pratiques sur bases de données de test : impact du floutage des QID s ur la mesure du k-anonymat (Dominique Blum) Questions des participants

Jeudi 30 mai 2019 – INS du Niger	
9h00-9h15	Installation des participants
9h15-10h50	Travaux pratiques sur bases de données de test : mesure de la l-diversité avec SPSS, Excel ou SQL (Dominique Blum) Questions des participants
10h50 -11h00	Pause
11h00-12h30	Travaux pratiques sur bases de données de test : impact du floutage sur la mesure de la l-diversité (Dominique Blum) Synthèse des applications pratiques (Dominique Blum) Questions des participants
12h30-14h00	Déjeuner / Pause
14h00-17h00	Techniques de réduction du risque de ré-identification (Dominique Blum) Questions des participants

Vendredi 31 mai 2019 – INS du Niger	
9h00-9h15	Installation des participants
9h15-10h50	Applications pratiques de réduction du risque de ré-identification (Dominique Blum) Questions des participants
10h50 -11h00	Pause
11h00-12h30	Questions posées par la diffusion d'une base de microdata (Dominique Blum) Questions des participants Quels outils pour en faire quoi : tableau sommaire (Dominique Blum) La juste place des outils « prêts à porter » (Dominique Blum) Questions des participants
12h30-14h00	Déjeuner / Pause
14h00-17h00	Récapitulatif des quatre journées (Dominique Blum) Recommandations générales (Dominique Blum) Synthèse des quatre journées – Tour de table général Questions ouvertes et discussion
17h00	Échange de coordonnées

4.3. Annexe 3 : Supports de formation



NiPN
National Information
Platforms for Nutrition

Bases de données compréhension, prévisions, perspectives et attentes

Guillaume POIREL
Analyste-Statisticien AT PNIN Niger

Abidjan, Côte d'Ivoire
Février 2019





Première journée

- Anonymisation, pseudonymisation
- Catégories de données d'une base de microdata
 - Identifiants nominatifs, données « sensibles », quasi-identifiants, autres données
- Ré-identification au sein d'une base de données pseudonymisée
- Mesures du risque de ré-identification au sein d'une base de données
 - estimation empirique
 - k-anonymat, l-diversité
 - autres méthodes





Quelques termes

- Micro-données (microdata en anglais)
 - base de données comportant des informations personnelles de granularité individuelle
- Données agrégées
 - tableau ou base de données comportant des informations relatives à des groupes d'au moins deux personnes
- Variables floutées
 - variables dont certaines modalités ont été constituées en plages de telle sorte que leur précision est réduite



NiPN
National Information
Platforms for Nutrition

Anonymisation, pseudonymisation





Anonymisation, *stricto sensu*

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	...
...





Anonymisation, *stricto sensu*

- Solution n°1 : disparition pure et simple du prénom et du nom

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	...
...



Anonymisation, *stricto sensu*

- Solution n°2 : pseudonymisation

« étiquette »	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
étiquette A	1952	Dandadji	Illéla	Tahoua	marié	5	...
étiquette B	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
étiquette C



Pseudonymisation : principe

- Remplacer les informations nominatives par une « étiquette à la fois unique et non identifiante » appelée pseudonyme
 - unique : deux enregistrements distincts doivent comporter deux étiquettes distinctes
 - non identifiante : le contenu de l'étiquette ne doit pas fournir de renseignement directement ou indirectement nominatif
- Conserver (**ou non**) séparément la correspondance entre l'étiquette et les informations nominatives qu'elle remplace
 - à décider selon l'objectif (cf. diapos suivantes)
 - s'inspirer de l'organisation de l'ABS (Australian bureau of statistics)



Pseudonymes : table de correspondance

- Données d'origine non anonymes

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	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL



Pseudonymes : table de correspondance

- Données d'origine non anonymes

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	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL



Pseudonymes : table de correspondance

- Données d'origine non anonymes

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	5	...
Brigi	RAFINI	07/04/1953	Iférouane	Arlit	Agadez	marié	5	...
Guillaume	POIREL



Pseudonymes : table de correspondance

- Données d'origine non anonymes

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	...

▶ Table de correspondance noms / pseudonymes + base de données pseudonymisées

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



Pseudonymisation : objectifs

- soit pour permettre de « remonter » au recueil d'origine si besoin
 - ⇒ conserver la correspondance « nom / pseudonyme » : **indispensable**
 - ⇒ usage **partagé** (producteur des données / service statistique)
- soit pour faciliter la manipulation des enregistrements
 - ⇒ conserver la correspondance : **totalemment inutile** (et même contre-productif)
 - ⇒ usage **interne** au service statistique



Pseudonymes à éviter, voire proscrire

- Initiales du prénom et du nom
 - MISSO-051, BRAFI-112, GPOIR-045, ...
- Code identifiant la région, la campagne de collecte, etc.
 - AGA2016-25087, MAR2017-11297, NIA2015-2564032, ...
- Attribution de numéros séquentiels
 - Ils peuvent révéler une information fondée sur un ordre naturel ou évident qui permet de « cibler » l'individu concerné :
 - tri alphabétique des noms
 - classement des villes et villages, départements, régions, etc.
 - déroulement chronologique de la collecte

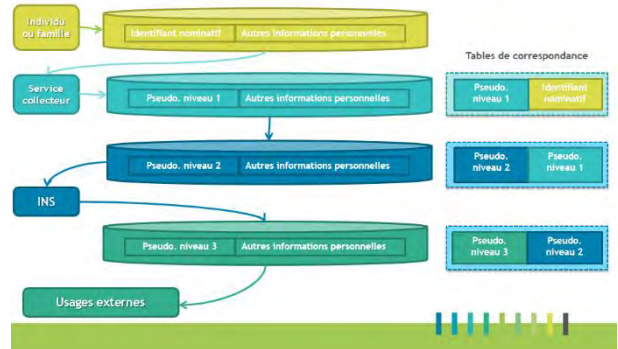


Comment pseudonymiser ?

- Trois méthodes distinctes
 - Numérotation séquentielle
 - Numérotation aléatoire
 - « Hachage cryptographique »
- Peuvent être mises en œuvre par le service collecteur des données
- Peuvent être combinées en cascade : étapes de pseudonymisation successives
 - pseudonymisation initiale par le collecteur des données
 - sur-pseudonymisation lors de la centralisation auprès du service statistique
 - sur-pseudonymisation lors de la diffusion d'une extraction de la base de données



Pseudonymisation « en cascade »



Pseudonymisation : méthode n° 1

- Numéro séquentiel
 - uniquement si les données ne sont pas fournies selon un tri préalable fondé sur un ordre naturel ou évident (ordre alphabétique des noms, classement habituel des villes et villages du département, des départements de la région, des régions du Niger, ordre chronologique de la collecte des données, etc.)
 - peut être mis en œuvre par le service collecteur des données

Prénom	NOM	pseudonyme	pseudonyme	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	00 000 001	00 000 001	1952	Dandadj	Iléla	Tahoua	marité polygame	5	...
Brigi	RAFINI	00 000 002	00 000 002	07/04/1953	Iférouane	Arit	Agadez	marité monogame	5	...
Guillaume	POIREL	00 000 003	France	marité monogame	3	...



Pseudonymisation : méthode n° 3 (1/2)

- « Hachage cryptographique »
 - indispensable uniquement dans le cas où le « chaînage » est nécessaire
 - chaînage chronologique : collectes itératives concernant les mêmes personnes
 - chaînage géographique : collectes concernant des personnes nomades
 - ne nécessite pas obligatoirement de conserver la table de correspondance !
 - en pratique on peut recourir aux fonctions de hachage publiques (MD5, SHA, etc.)
 - peut être mis à la disposition du service collecteur des données



Pseudonymisation : méthode n° 2

- Numéro aléatoire
 - assez simple à développer et mettre en œuvre
 - peut être mis à la disposition du service collecteur des données

Prénom	NOM	pseudonyme	pseudonyme	date de naissance	lieu de naissance	département de naissance	région de naissance	situation matrimoniale	nombre d'enfants	...
Mahamadou	ISSOUFOU	870 145 032	870 145 032	1952	Dandadj	Iléla	Tahoua	marité polygame	5	...
Brigi	RAFINI	802 237 751	802 237 751	07/04/1953	Iférouane	Arit	Agadez	marité monogame	5	...
Guillaume	POIREL	537 501 223	537 501 223	France	marité monogame	3	...



Pseudonymisation : méthode n° 3 (2/2)

- « Message en entrée » constitué des informations nominatives
 - ex: **1530499337175070419531**
- Application de la fonction de hachage sur le « message »
- « Clef de hachage » en sortie
 - ex: **9p282kPH165bnUj3h075MnAx22yTM113FdF4n63967mAA83TtZ8582qe2wS045**
 - longueur de la clef « suffisamment longue »
 - reproductibilité parfaite (le même message produit toujours la même clef)
 - taux de collision quasiment nul (deux messages distincts ne peuvent produire la même clef)
 - effet avalanche maximum (un changement infime du message produit une clef totalement différente)





Conserver la correspondance

Pourquoi conserver la table de correspondance ?

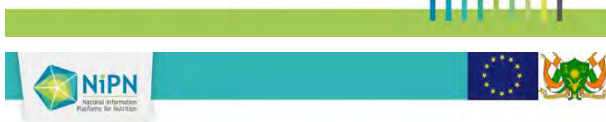
- pour « remonter » de l'enregistrement pseudonymisé au recueil d'origine
- pour « redescendre » du recueil d'origine à l'enregistrement pseudonymisé
 - sauf dans le cas du hachage cryptographique

Qui doit conserver la table de correspondance ?

- l'organisme qui est « en amont »

Comment conserver la table de correspondance ?

- cloisonnement



Identifiants (in)directement nominatifs

- Directement nominatifs
 - Prénom et NOM
- Indirectement nominatifs, par utilisation de registres ou d'annuaires
 - Adresse complète de résidence
 - Numéro de passeport, de carte nationale d'identité, de registre d'état-civil
 - Numéro de sécurité sociale, de permis de conduire
 - Numéro de téléphone
 - Adresse de courriel
- Indirectement nominatifs, par croisement de bases et de registres
 - Longitude + latitude du lieu de résidence => adresse complète de résidence
 - Adresse IP (internet protocol) => adresse complète de résidence



Quasi-identifiants (QID)

- Tentative de définition
 - Information plus ou moins précise qu'un « attaquant potentiel » peut détenir « assez couramment » au sujet d'un individu, mais ne permettant pas à elle seule de l'identifier au sein de la base de données
- Risques liés aux QID
 - Ré-identification par unicité des combinaisons de QID
 - Ré-identification par croisement avec d'autres sources



Données « sensibles »

- Données personnelles et confidentielles dont la révélation est une atteinte à la vie privée de la personne concernée
 - Source et montant des revenus
 - Descriptif et montant du patrimoine
 - Décisions administratives privées
 - Décisions judiciaires privées
 - État de santé (diagnostics, actes chirurgicaux, médicaments, ...)
 - ...
- La protection de la confidentialité des données sensibles est la seule justification de l'anonymisation



Exemples de QID non spécifiques

- QID les plus courants (toutes bases de données)
 - Date de naissance, âge
 - Sexe
 - Village, département, région de résidence
 - État marital, nombre d'enfants
 - Profession, catégorie socio-professionnelle
 - ...





Exemples de QID spécifiques

- **Base de données médico-administratives hospitalières (française)**
 - Date d'entrée, date de sortie, durée de séjour
 - Mode d'entrée, mode de sortie
 - Identification de l'établissement d'hospitalisation
- 2009 : 17 millions de séjours pseudonymisés
- 89% de ces séjours présentent des combinaisons de QID uniques
 - Accès aux données de santé de tous les individus hospitalisés en France



Exemples de QID spécifiques

- **Travaux de recherche américains**
 - Date de naissance complète
 - Sexe
 - Code postal de résidence complet
- 87% des résidents américains présentent des combinaisons uniques de ces trois QID



Qu'entend-on par « ré-identification » ?

- Retrouver au sein d'une base de données les enregistrements relatifs à une personne dont on connaît déjà certains QID
- Retrouver des combinaisons de QID dont on dispose dans d'autres bases de données, qui fournissent l'identité des personnes concernées
 - cf. 1997 : Latanya Sweeney ré-identifie William Weld



Qu'entend-on par « ré-identification » ?

- **Principe**
 - fondé sur l'unicité (ou la rareté) de la combinaison des QID relatifs à la personne ciblée (ou au ménage ciblé).
- **Objectifs (rarement licites)**
 - lever la confidentialité sur la personne concernée
 - lever la confidentialité sur certains attributs relatifs à la personne concernée
 - les deux à la fois
- **Ne concerne pas nécessairement une célébrité**



Anonymisation versus ré-identification

- **Anonymisation (ou plutôt pseudonymisation)**
 - consiste à faire disparaître des informations inutiles pour le travail statistique proprement dit
 - processus technique relativement simple à maîtriser
- **Empêcher la ré-identification**
 - consiste à réduire, voire supprimer le risque lié à la combinaison des QID
 - sans supprimer les QID, indispensables au travail statistique proprement dit
 - processus technique un peu plus complexe à maîtriser



Y a-t-il un risque de ré-identification ?

- Dans « notre » base de données
 - Établir la liste exhaustive des QID
 - Commencer par une estimation empirique
- 1^{er} cas : risque empirique faible
 - ⇒ résoudre la question des cas extrêmes (*outliers*)
- 2^e cas : risque empirique moyen ou fort
 - ⇒ calculer le risque réel
 - ⇒ se fixer des objectifs (seuils ou plafonds)
 - ⇒ résoudre une par une toutes les situations hors des limites

Mesures du risque de ré-identification au sein d'une base de données

- méthode empirique
- *k*-anonymat, *l*-diversité
- autres méthodes



k-anonymat (*k*-anonymity)

- Dans une base de données
 - on distingue toutes les combinaisons effectives des QID
 - on observe le nombre d'enregistrements relevant de chacune d'elles
 - la valeur la plus faible est appelé *k*-anonymat
- Les « normes habituelles »
 - *k*-anonymat inférieur à 5 : mauvais
 - *k*-anonymat compris entre 5 et 9 : médiocre
 - *k*-anonymat supérieur ou égal à 10 : satisfaisant
 - *k*-anonymat supérieur à 30 : très satisfaisant



Interprétation du *k*-anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes
donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage



Cas particulier : l'échantillonnage

- Moins la base de données est exhaustive, moins elle présente de risque de ré-identification
 - il s'agit de l'exhaustivité par rapport à son propre périmètre
 - sous réserve de ne pas révéler certaines caractéristiques de la population dont l'échantillon est issu
 - sous réserve de ne pas pouvoir la recouper avec une autre base qui serait exhaustive et comporterait les mêmes QID
 - sous réserve de ne pas révéler qui fait partie de l'échantillon



Estimation empirique : méthode

- Pour chaque QID
 1. classer les modalités du QID par « part relative » décroissante
 - soit sur la base du dénombrement réel
 - soit à dire d'expert (distributions connues pour être uniforme ou non, etc.)
 2. déterminer le nombre de modalités qui à elles seules totalisent au moins 80% de l'effectif total
- Pour l'ensemble de la base
 - calculer le produit de toutes les valeurs obtenues à l'étape 2 précédente
 - ce produit est l'estimation empirique du nombre de combinaisons de QID distinctes
 - diviser l'effectif total de la base par ce produit
 - au dessus de 30 le risque est faible
 - en dessous de 10 le risque est fort
 - entre 10 et 30 le risque est moyen



Intérêt du *k*-anonymat

- Que représente-t-il ?
 - L'effectif de la combinaison de QID la plus rare
 - Un indicateur de la rareté des combinaisons de QID de toute la base
- On peut le compléter par
 - Nombre total de combinaisons de QID distinctes dans la base de données
 - Nombre de combinaisons de QID ayant pour effectif la valeur du *k*-anonymat
 - Effectif moyen des combinaisons de QID
- A quoi sert-il ?
 - C'est une métrique qui participe à la qualification du risque relatif à une base
 - Permet de « monitorer » l'évolution du risque de ré-identification obtenue par les mesures que l'on va mettre en œuvre pour réduire ce risque



Interprétation du *k*-anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes
donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage





Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes

donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes

donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

donc

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage



Interprétation du k -anonymat > 10

Aucun individu ne se trouve dans un groupe de moins de 10 personnes

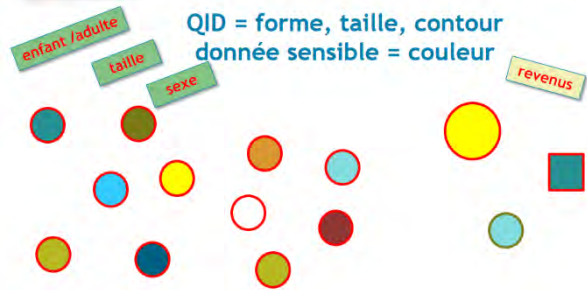
donc

Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires

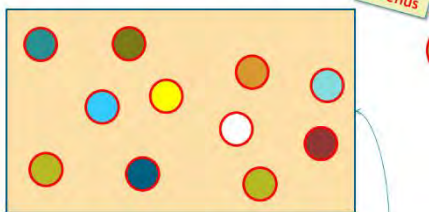
donc

Il est impossible pour un attaquant de tirer des conclusions relatives à un individu ou un ménage à partir des données sensibles de la base de données, en se fondant sur les QID dont il dispose relatifs à cet individu ou ce ménage

Sauf que ce n'est pas tout à fait suffisant !



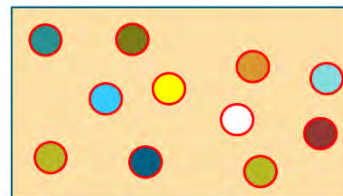
QID = forme, taille, contour
 donnée sensible = couleur



rond, petit, contour rouge



QID = forme, taille, contour
 donnée sensible = couleur



Il est impossible de lever la confidentialité des données sensibles relatives à tel ou tel membre du groupe.

L'hétérogénéité des couleurs le prouve.

rond, petit, contour rouge



NiPN National Information Platform for Fisheries

QID = forme, taille, contour
donnée sensible = couleur

rond, petit, contour rouge

NiPN National Information Platform for Fisheries

QID = forme, taille, contour
donnée sensible = couleur

rond, petit, contour rouge

Dans ce cas limite, peu importe de savoir quel enregistrement correspond à quel individu : il est évident que tous les individus du groupe présentent la même information sensible.

Le risque de voir levée la confidentialité des données sensibles relatives à tel ou tel membre du groupe est élevé.

NiPN National Information Platform for Fisheries

QID = forme, taille, contour
donnée sensible = couleur

rond, petit, contour rouge

NiPN National Information Platform for Fisheries

***l*-diversité (*l*-diversity)**

- Dans une base de données
 - on distingue toutes les combinaisons effectives des QID
 - on observe le nombre de modalités distinctes des données sensibles que présentent chacune d'elles, qu'on appelle sa diversité
 - la valeur la plus faible est appelé *l*-diversité
- Les « normes habituelles »
 - *l*-diversité inférieure à 3 : mauvais
 - *l*-diversité supérieure ou égale à 3 : satisfaisant
 - *l*-diversité supérieure ou égale à 5 : très satisfaisant

NiPN National Information Platform for Fisheries

QID = forme, taille, contour
donnée sensible = couleur

rond, petit, contour rouge

Nombre de modalités distinctes des données sensibles (couleur)
10

NiPN National Information Platform for Fisheries

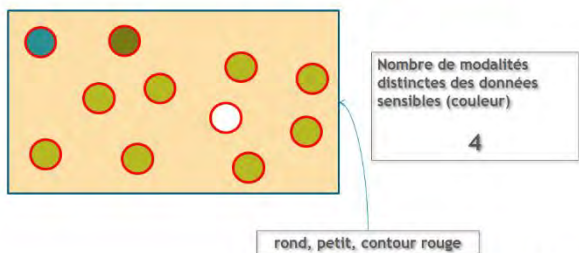
QID = forme, taille, contour
donnée sensible = couleur

rond, petit, contour rouge

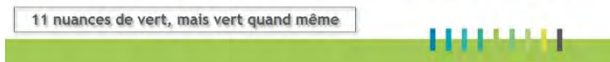
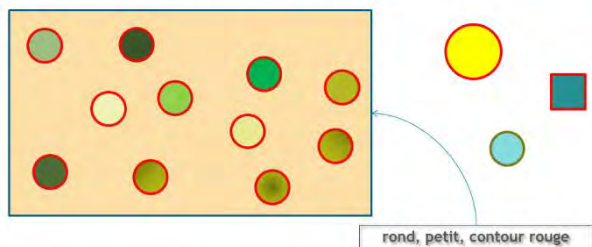
Nombre de modalités distinctes des données sensibles (couleur)
1



QID = forme, taille, contour donnée sensible = couleur



I-diversité : une difficulté pratique



I-diversité : autre difficulté pratique

• Il y a souvent plusieurs données sensibles : de laquelle doit-on mesurer la I-diversité ?

- on retient la I-diversité de la variable sensible qui est la moins diverse
 - soit elle est bien connue des gens du métier ou des experts et alors on ne calcule la I-diversité que pour cette variable
 - soit on hésite entre plusieurs variables sensibles, et dans ce cas on calcule la I-diversité pour chacune de ces variables pour ne retenir finalement que la plus faible



Intérêt de la I-diversité

- **Que représente-t-elle ?**
 - L'hétérogénéité des données sensibles de la combinaison de QID la plus uniforme
 - Un indicateur de l'hétérogénéité au sein de toutes les combinaisons de QID
- **On doit la compléter par**
 - Proportion de combinaisons ayant une diversité égale à 1
 - Proportion de combinaisons ayant une diversité égale à 2
 - Proportion de combinaisons ayant une diversité de 3 ou plus
- **A quoi sert-elle ?**
 - C'est une métrique qui participe à la qualification du risque relatif à une base
 - Permet de « monitorer » l'évolution du risque de ré-identification obtenue par les mesures que l'on va mettre en œuvre pour réduire ce risque



I-diversité : une difficulté pratique

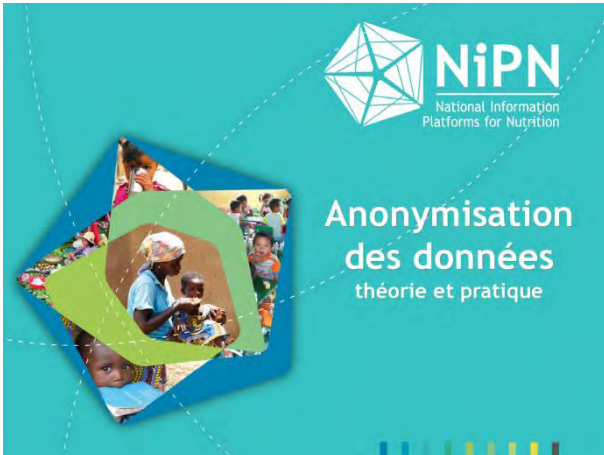
- **Les modalités précises des données sensibles peuvent être distinctes mais très proches**
 - un attaquant pourrait finalement tirer des conclusions « assez précises » sur tous les individus qui composent le groupe
 - la confidentialité des informations relatives à tous les individus du groupe est compromise, même si l'on ne peut déterminer qui est précisément chaque individu du groupe
 - pour le calcul de la diversité, il serait justifié d'utiliser une variable moins précise, de sorte que toutes ces modalités n'en fassent plus qu'une seule
 - pour améliorer la protection des données sensibles, il faudrait renforcer la diversité du groupe, par exemple en augmentant sa taille (fusion avec un autre groupe, dont les QID seraient proches)



Autre méthode de calcul du risque de ré-identification

- pour une combinaison donnée des QIDs
 - nombre d'individus présentant cette combinaison dans la population : \mathcal{N}
 - nombre d'individus présentant cette combinaison dans l'échantillon : \mathcal{K}
- **risque de ré-identification pour cette combinaison**
 - si l'on sait que l'individu recherché se trouve dans l'échantillon : $1 / \mathcal{K}$
 - si on l'ignore : $\mathcal{K} / \mathcal{N}$
- **remarques**
 - calculer un risque pour chacune des combinaisons
 - la plupart du temps on connaît \mathcal{K} mais pas \mathcal{N}
 - il faut estimer \mathcal{N} à partir de \mathcal{K}
 - mais en général, si le facteur d'échantillonnage est \mathcal{E} , le risque est dans l'intervalle $[\mathcal{E} / 2, \mathcal{E} * 2,5]$
 - exemple, pour un échantillonnage au 100^{ème} le risque est compris entre 1/200 et 1/40
 - ne dispense pas de se préoccuper de la I-diversité !





Résumé de l'épisode précédent

1. suppression du risque d'identification immédiat
 - suppression des identifiants individuels directs et/ou indirects
 - pseudonymisation (le plus souvent)
2. contrôle du risque de « ré-identification »
 - évaluation du risque
 - estimation empirique
 - métrique habituelle : k -anonymat et k -diversité
 - déterminer quelle stratégie de réduction du risque mettre en œuvre



Hachage cryptographique avec Excel (2/2)



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



Calcul du k -anonymat avec SQL (2/2)

Pseudonyme	région		sexe		âge		situation matrimoniale		dépenses de santé		dépenses de santé		ressources	
	QID n°1	QID n°2	QID n°3	...	QID n°x	donnée sensible n°1	donnée sensible n°2	...	donnée sensible n°y
1	q1	q2	q3	...	qx	v1	v2	...	vy

```
SQL> select q1, q2, q3, [...], qx, count(*) as k
from [nom_de_la_table]
group by q1, q2, q3, [...], qx
```



Deuxième journée

- Travaux pratiques



Hachage cryptographique avec Excel (1/2)

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

Function ComputeHash(fid)
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(fid))))
End Function
```



Calcul du k -anonymat avec SQL (1/2)

Pseudonyme	région		sexe		âge		situation matrimoniale		dépenses de santé		dépenses de santé		ressources	
	QID n°1	QID n°2	QID n°3	...	QID n°x	donnée sensible n°1	donnée sensible n°2	...	donnée sensible n°y
1	q1	q2	q3	...	qx	v1	v2	...	vy



Réduction des risques de ré-identification et de divulgation

Actions sur les QID
Actions sur les données sensibles



Troisième journée

• Méthodes pour limiter le risque de ré-identification

1. intervenir sur les QIDs
 - suppression des quasi-identifiants (QIDs) inutiles
 - « floutage » des QIDs utiles (sans réduire leur « utilisabilité »)
2. intervenir sur les données sensibles
 - il ne s'agit plus à proprement parler de techniques d'anonymisation
 - trois types de méthodes



Autrement dit

- **Base nominative de données brutes**
 - suppression des informations directement ou indirectement nominatives
 - introduction de pseudonymes
- **Base pseudonymisée de données brutes**
 - évaluer empiriquement le risque de ré-identification par les combinaisons de QID
 - calculs exacts du k -anonymat et de la l -diversité
- **Et si k -anonymat < 10 ou l -diversité < 3**
 - réduire le risque en réduisant le nombre de combinaisons (jouer sur les QIDs)
 - réduire le risque de divulgation en réduisant la précision des données sensibles



Réduction du risque de ré-identification

Intervenir sur les QIDs

Les méthodes

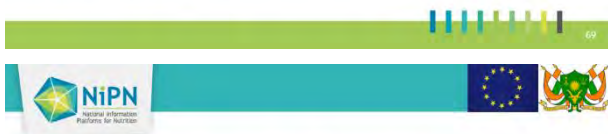
- méthodes non perturbatrices
 - « floutage » des QIDs (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des QIDs
- méthodes perturbatrices
 - bruitage des QIDs
 - post-randomisation des QIDs
 - permutation des valeurs de QIDs
- création de jeux de données virtuelles



Action sur les QIDs : « floutage »

(1/2)

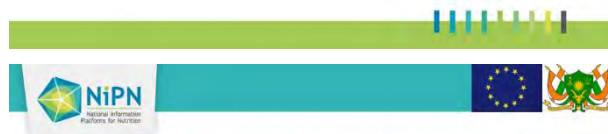
- **Terminologie officielle : « généralisation »**
 - réduire la précision des QID en regroupant plusieurs modalités
 - QID quantitatifs : plages de valeurs (âge, ressources, dépenses, etc.)
 - QID catégoriels ou qualitatifs : catégories plus larges
 - commencer par flouter les QID ayant le plus grand nombre de modalités
 - localisations géographiques (directes ou indirectes)
 - dates
 - référentiels détaillés (professions, etc.)
 - âge



Action sur les QIDs : « floutage »

(2/2)

- **Garder à l'esprit qu'en général**
 - réduire les modalités d'un facteur de 10 réduit le nombre de groupes d'un facteur de 2, voire 3 mais pas plus
 - réduire les modalités des QID
 - agit très peu sur l'effectif de la combinaison de QID la plus rare (le k -anonymat)
 - mais agit nettement sur les effectifs des combinaisons les plus denses
- **Conséquence**
 - plutôt qu'une généralisation globale, mettre en œuvre une généralisation locale (ou « différenciée »)
 - avec un inconvénient : les modalités sont inhomogènes



Réduction du risque de ré-identification

Intervenir sur les QIDs

Les méthodes

- méthodes non perturbatrices
 - « floutage » des QIDs (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des QIDs
- méthodes perturbatrices
 - bruitage des QIDs
 - post-randomisation des QIDs
 - permutation des valeurs de QIDs
- création de jeux de données virtuelles





Perturbation des QIDs

- **Création de bruit**
 - ajout de valeurs aléatoires (généralement faibles) aux QIDs quantitatifs
 - sous réserve de préserver les moyennes, variances, corrélations...
- **Permuter les valeurs de QIDs de deux enregistrements**
 - sous réserve de réciprocité pour ne pas biaiser l'ensemble
 - sous réserve de ne pas créer des absurdités détectables



« Masquage » des données

- **Principe**
 - modifier les données sensibles de certains enregistrements
 - tout en conservant « certaines » de leurs caractéristiques statistiques
 - moyennes, corrélations, etc.



« détails » à prendre en considération

- **Échantillonnage**
 - attention au croisement avec une base de données externe
- **Suppression d'enregistrements**
 - sous réserve d'en ajuster l'impact
- **Traitement des *outliers* (valeurs extrêmes)**
 - limite inférieure
 - limite supérieure
- **Suppression d'une donnée trop discriminante**
 - transformée en valeur manquante
 - sous réserve de ne pas communiquer sur ce « cas exceptionnel »



Résumé de l'épisode précédent

1. **interventions sur les QIDs**
 - sélection itérative du « meilleur » QID à re-traiter
 - suppression de QIDs inutiles
 - « floutage » de certaines modalités de QIDs
 - calcul itératif du k-anonymat sous Excel
2. **interventions sur les données sensibles**
 - « floutage » de certaines modalités de données sensibles
 - calcul itératif de la l-diversité sous Excel



Réduction du risque de divulgation de données sensibles (mais ce n'est plus de l'anonymisation)

Les méthodes (= masquage des données)

- **méthodes non perturbatrices**
 - « floutage » des données sensibles (= recodage en plages de valeurs)
 - suppression locale de certaines valeurs des données sensibles
- **méthodes perturbatrices**
 - bruitage des données sensibles
 - micro-agrégation des données sensibles
- **création de jeux de données virtuelles**

Dans tous les cas, il faut ensuite évaluer la perte d'information



Perturbation de l'information

- **Création de bruit**
 - ajout de valeurs aléatoires (généralement faibles) aux valeurs recueillies
 - sous réserve de préserver les moyennes, variances, corrélations...
- **Permuter certaines données sensibles de deux enregistrements**
 - sous réserve de réciprocité pour ne pas biaiser l'ensemble
 - sous réserve de ne pas créer des absurdités détectables
- **Micro-agrégation**
 - on remplace les données par une « moyenne locale »



NiPN
National Information
Platforms for Nutrition

**Anonymisation
des données**
théorie et pratique



Quatrième journée

- **Essais manuels de méthodes perturbatrices**
 1. **sur les QIDs**
 - bruitage
 - permutations
 - calculs itératifs du k-anonymat et de la l-diversité
 2. **sur les données sensibles**
 - bruitage
 - micro-agrégation
 - calculs itératifs de la l-diversité





Recommandations

1. Ne pas réduire le sujet à la question de la pseudonymisation
2. Ne pas espérer le logiciel-miracle
3. Fixer les objectifs en fonction du type de base de données et du type de diffusion
4. L'échantillonnage peut rassurer à tort
5. Huit à dix QJD au maximum
6. Attention aux pseudo-QJD qui sont en fait des vrais identifiants indirectement nominatifs
7. Maîtriser l'approche empirique pour le calcul du risque de ré-identification
8. Se concentrer sur les combinaisons de QJD de faible effectif
9. Privilégier la « généralisation locale » des QJD
10. Recourir à la « généralisation locale » des données sensibles en cas de besoin



Vos questions et commentaires



Encore des recommandations

- Garder à l'esprit : la diffusion en *open data* est en général « *one shot* »
- Maîtriser le processus de A à Z
- Prévoir trois mois à temps plein pour une base de 150 à 200 variables dont 5 à 10 QJD
- Questions ciblées par courriel : dblum@le-pmsi.fr



Merci pour votre attention,
et bonnes pseudonymisations
non ré-identifiantes !



NiPN
National Information Platforms for Nutrition

Anonymisation des données
théorie et pratique



Démographie du Niger

	1977	1988	2001	2012	2013	2014	2015	2016	2017	2018
Niger	5 102 990	7 251 626	11 060 291	16 993 563	17 679 760	18 389 164	19 124 883	19 865 067	20 651 070	21 466 863
Agadez	134 985	208 828	321 639	478 833	495 311	512 148	529 658	547 755	566 447	585 737
Diffa	167 389	189 091	346 595	586 648	607 732	627 529	648 049	669 307	691 356	714 242
Dosso	693 207	1 018 895	1 505 864	2 034 324	2 113 733	2 195 788	2 280 703	2 368 651	2 459 812	2 554 379
Maradi	949 747	1 389 433	2 235 748	3 365 969	3 511 327	3 663 102	3 821 593	3 987 165	4 160 231	4 340 983
Tahoua	993 615	1 308 598	1 972 729	3 304 193	3 432 320	3 564 239	3 699 907	3 839 457	3 983 172	4 133 384
Tillabéri	928 849	1 328 283	1 889 515	2 701 408	2 814 086	2 930 976	3 052 368	3 175 731	3 280 333	3 409 676
Zinder	1 002 225	1 411 061	2 080 250	3 506 758	3 653 746	3 806 825	3 966 348	4 132 321	4 305 953	4 487 009
Niamey	242 973	397 437	707 951	1 015 430	1 051 605	1 088 557	1 126 257	1 164 680	1 203 766	1 243 453

