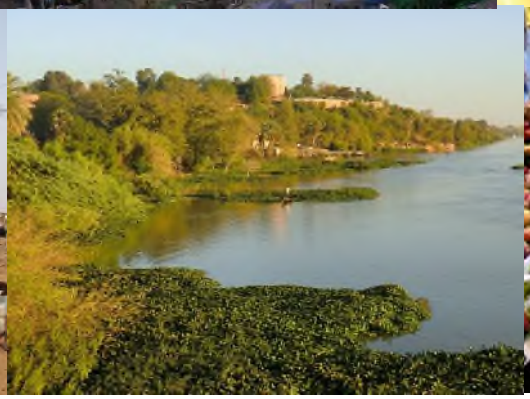
NIGER UNION EUROPEENNE




SOFRECO

RAPPORT DE MISSION

FORMATION A L'ANONYMISATION DES DONNEES AUPRES DE L'INSTITUT NATIONAL DE LA STATISTIQUE DU NIGER ET DES DIRECTIONS STATISTIQUES DES MINISTERES DE LA SANTE ET DE L'EDUCATION

Assistance Technique pour la Mise en Place de la Plateforme Nationale d'Information pour la Nutrition au Niger

Juin 2018

NiPN
Plateformes nationales d'information pour la nutrition

TABLE DES MATIERES

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) », 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 et les parties prenantes 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 (créer au sein de l'INS une unité de mission capable de gérer, d'analyser et de diffuser l'information relative à la nutrition) et plus particulièrement dans le souci de pouvoir collecter et traiter les données des Directions Statistiques des Ministères clés, l'AT a prévu la mobilisation d'une Expertise Court Terme pour une formation en anonymisation des données.

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.

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 inexistants (seuls les annuaires sont transmis sous format 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(SSN)). 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 apparait 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 doit se dérouler en deux sessions de formation de cinq jours. La première phase, qui fait l'objet de ce rapport, s'est déroulée du 7 au 11 mai 2018 à Niamey. Outre des représentants de l'INS, elle a concerné des représentants du HC3N, du Ministère de l'éducation et du Ministère de la santé, soit au total dix-neuf personnes. La liste des participants est donnée en annexe 1.

La seconde phase se déroulera également à Niamey, en février 2019. Cette seconde formation devrait concerner d'autres participants provenant des autres Ministères concernés.

Une part de la formation est consacrée aux aspects théoriques, dont certains sont ensuite mis en application lors d'exercices pratiques. Pour pouvoir en profiter pleinement, il est nécessaire que chaque participant dispose d'un ordinateur portable configuré avec son outil familier de traitement statistique (en pratique SPSS ou Stata). Dans la mesure du possible, les participants doivent disposer 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 sont 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 et le support principal de la formation en annexe 3.

2.1 Première journée

Après une rapide présentation de ses attentes par chaque participant, le chef de 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. En effet, il ne suffit pas de faire disparaître le nom et le prénom pour qu'une base de données devienne véritablement anonyme. Les participants ont pu réaliser que ce que l'on a pour habitude d'appeler « base de données anonymisée » est en réalité une « base de

données pseudonymisée ». Le pseudonyme est en réalité un identifiant personnel indirect.

Les clefs et principes d'une bonne pseudonymisation sont présentés. L'après-midi est ensuite consacrée à 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.

Au terme de la journée, les participants s'étant approprié tous ces concepts réalisent pourquoi la formation ne s'achève 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, elle permet encore dans la plupart des cas de ré-identifier les individus qui la composent.

2.2 Deuxième journée

Au cours de la matinée, le risque de ré-identification évoqué la veille est approfondi et décortiqué, 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.

Une partie de l'après-midi est consacrée à la présentation et à la mise en œuvre pratique 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.

Le reste de l'après-midi est consacré à la présentation des deux métriques les plus utilisées pour rendre compte du risque de ré-identification d'une base de données : le k-anonymat et la l-diversité.

2.3 Troisième journée

Une bonne partie de la matinée est consacrée à l'installation et à la configuration des postes individuels des participants. Le reste de la matinée et toute l'après-midi sont consacrés à des travaux pratiques sur des bases de données de tests :

- Mesure empirique du risque de ré-identification ;
- Mesure du k-anonymat ;
- Floutage des QID et impact sur le k-anonymat.

2.4 Quatrième journée

La matinée est entièrement consacrée à la poursuite des travaux pratiques sur les bases de données de test :

- Mesure de la l-diversité ;
- Floutage des QID et impact sur la mesure de la l-diversité.

Pendant l'après-midi, après une synthèse des travaux pratiques, les techniques de réduction du risque de ré-identification par masquage des données sensibles sont passées en revue : suppression partielle d'information (échantillonnage, suppression d'enregistrements, floutage des données sensibles, traitement des

outliers (valeurs aberrantes ou valeur exceptionnelle), suppression des valeurs discriminantes) et perturbation de l'information (création de bruit, permutation de données, micro-agrégation).

2.5 Cinquième journée

Une partie de la matinée est consacrée à divers exercices mettant en pratique les techniques de réduction du risque de ré-identification évoquées la veille et permettant d'observer leur effet réel avec les bases de données de test.

Sont ensuite abordées les questions posées par la diffusion d'une base de données personnelles.

L'après-midi conclut la formation 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 voulue délibérément pour 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 autonomes 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

En revanche deux points pourraient être améliorés lors de la prochaine session :

- La préparation des postes de travail individuels : beaucoup de temps a été consacré à la préparation matérielle et logicielle, à l'installation des outils et à leur paramétrage initial. Cela devrait être réalisé en amont, car il ne s'agit pas d'outils spécifiques, mais précisément au contraire des outils dont les participants doivent être familiers ;
- La préparation de jeux de données de tests : chaque organisme participant (chaque ministère) est confronté à des problématiques spécifiques, et seules des données issues de son activité habituelle permettent 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 est indispensable que

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

3.1.3 Suggestion pour février 2019

Sous réserve d'une préparation plus efficace des participants (cf. point précédent), il convient d'envisager la réalisation de 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.

Cette troisième journée consacrée par les participants à leurs exercices pratiques « personnalisés » serait donc libérée pour le formateur. Ainsi, lors de la seconde session de formation (février 2019), ce troisième jour pourrait permettre de réunir les participants de la première formation (mai 2018) afin de répondre aux questions qu'ils n'auront pas manqué de se poser lors de la mise en œuvre des techniques d'anonymisation dans leurs services respectifs et d'échanger de manière très pragmatique sur leurs expériences individuelles.

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 connaissent bien la question de l'anonymisation des données et sont soumis à une forte pression pour produire des « bases de données anonymes ». La plupart des participants s'attendaient 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'ambigüité du terme « anonyme » : au sens propre, est anonyme une base de données dans laquelle les identifiants nominatifs n'apparaissent pas, ni directement ni indirectement. Lorsque les identifiants nominatifs sont remplacés par des pseudonymes numériques séquentiels, numériques aléatoires ou résultant d'un hachage cryptographique, il n'y a pas besoins de beaucoup de moyens informatiques. Mais, en pratique, supprimer les identifiants ne rend pas anonyme les individus, 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 » puisse résoudre en quelques instants les problématiques d'anonymat des données doit être combattue et abandonnée. La preuve a été donnée par le Chef de mission de l'AT PNIN qui a montré avec l'aide du Formateur que les bases des Enquêtes Démographique et de Santé (EDS) de l'INS n'étaient pas anonymes et qu'il était possible d'identifier des individus à travers les QID, bases considérées par l'INS comme répondant aux critères d'anonymisation.

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. 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 l'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 intégrale particulièrement difficile, voire impossible ;
- D'autre part le nombre de variables sensibles : les bases de données étudiées pendant la formation (base de l'enquête conjointe sur la vulnérabilité à l'insécurité alimentaire des ménages de 2014) comportent de 250 à 300 variables sensibles distinctes, sur lesquelles il faut 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* (valeurs aberrantes et/ou valeurs exceptionnelles);
- 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

- Un échantillon avec un taux d'échantillonnage trop élevé (supérieur à 1/20) ;
- Ou un 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* (valeurs aberrantes et/ou valeurs exceptionnelles);
- Monitorage 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 (07-11/05/2018)

	Prénoms et Noms	*Structure*	*Téléphone*	*Email*
1	Abdoul Karim Harouna Issa	INS/ENSPS	96 74 19 94	habdoulkarim@ins.ne
2	Aïtchedji Julinenne	DI/INS	96 96 07 06	jaitchedji@ins.ne
3	Ali Ousmane	INS	96 21 44 13	aousmane@ins.ne
4	Clémence Rieuneau	AT/HC3N	89 30 29 15	clemence.rieuneau@giz.de
5	Dabal Kadiatou	INS	98 60 88 25	kdabal@ins.ne
6	Guillaume POIREL	AT PNIN INS/HC3N	90 01 67 48	guillaumepoirel@gmail.com guillaume.poirel@sofreco.biz gpoirel@ins.ne
7	Habiba Aminato	DI/INS	96 55 78 97	Haminato@ins.ne
8	Hamadou Seyni	INS	99 75 90 09	aseyni@ins.ne
9	Hassane Amina	DEP/MEP	97 88 88 63	hima.amina@yahoo.fr
10	Ibrahim Lemane	HC3N	96 96 28 69	ilemane@387@gmail.com
11	Ibrahima Salamatou	DER/INS	96 43 05 14	sibrahim@ins.ne
12	Issaka Karimouna	INS	96 97 44 51	kissaka@ins.ne
13	Issiak Balarabé Mahamane	Coordonnateur PNIN/INS	99 75 91 20 94 63 60 30	mbalarabe@ins.ne mahamane.issiak@yahoo.fr
14	Maman Lawaly Boukari	DI/INS	96 32 33 02	lboukari@ins.ne
15	Mouctar Mamoudou	DS/MEP	98 15 85 55	mammouctar@yahoo.fr
16	Oumarou Sanifatou	DS/MSP	96 46 09 44	sanifaadamou@gmail.com
17	Zabeirou Sabiou	DEP/MSP	96 55 40 87	sabiouz@yahoo.fr
18	Tie Kasson Marx	AT PNIN INS/HC3N	80 63 59 98	kasson/marx.tie@sofreco.biz
19	Fatoumata Lankoande	AT PNIN INS/HC3N	98 25 52 37	fatoumata/lankoande@sofreco.biz

4.2 Annexe 2 : Chronogramme de la formation « Anonymisation des données » 05-09/05/2018

Chronogramme de la formation Anonymisation des données

Plateforme Nationale d'Information pour la Nutrition (PNIN) au Niger

Lundi 5 Mai 2018 – Hôtel Terminus	
09h30 - 09h45	**Installation des participants**
09h45 - 10h00	Mot de bienvenue de Monsieur Poirel Guillaume (Chef de mission AT PNIN) et de Monsieur Blum Dominique (Expert en anonymisation des données)
10h00 - 10h20	**Pause-café**
10h20 - 12h30	**Préambule : objectifs du portail, attentes et fonctionnement** **Présentation et définitions : anonymisation, pseudonymisation** (Blum Dominique) Questions des participants
12h30 - 14h00	**Déjeuner**
14h00 - 16h30	**Les catégories de données d'une base de microdata : identification nominative directe ou indirecte – Quasi-identifiants – Données sensibles** (Blum Dominique)
16h30 - 17h00	Questions des participants **Pause-café**

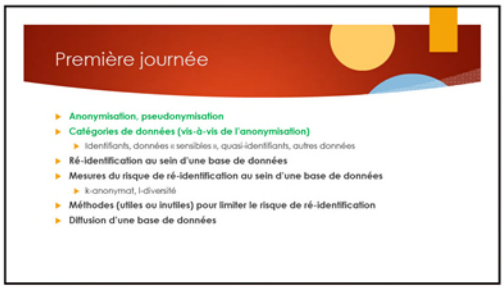
Mardi 6 Mai 2018 – Hôtel Terminus	
09h00 - 09h15	**Installation des participants**
09h15 - 10h00	**Le concept de ré-identification d'une base pseudonymisée** (Blum Dominique) Questions des participants
10h00 - 10h20	**Pause-café**
10h20 - 12h30	**Ré-identification par les QID** (Blum Dominique) Questions des participants
12h30 - 14h00	**Déjeuner**
14h00 - 16h30	**Mesure empirique du risque de ré-identification** **Les deux mesures : k-anonymat et l-diversité** (Blum Dominique)
16h30 - 17h00	Questions des participants **Pause-café**

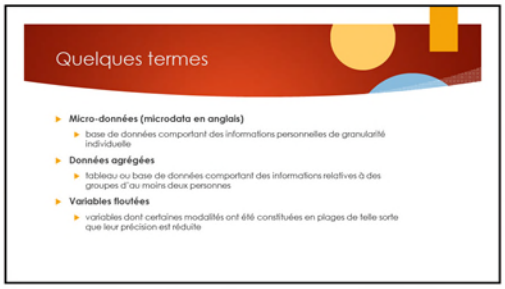
Mercredi 7 Mai 2018 – Hôtel Terminus	
09h00 - 09h15	**Installation des participants**
09h15 - 10h00	**Travaux pratiques sur bases de données de test : mesure empirique du risque de ré-identification** (Blum Dominique) Questions des participants
10h00 - 10h20	**Pause-café**
10h20 - 12h30	**Travaux pratiques sur bases de données de test : mesure du k-anonymat avec SPSS, Excel ou SQL** (Blum Dominique) Questions des participants
12h30 - 14h00	**Déjeuner**
14h00 - 16h30	**Travaux pratiques sur bases de données de test : impact du floutage des QID sur la mesure du k-anonymat** (Blum Dominique)
16h30 - 17h00	Questions des participants **Pause-café**

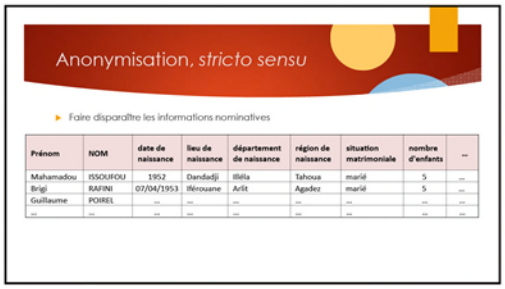
Jeudi 8 Mai 2018 – Hôtel Terminus	
09h00 - 09h15	**Installation des participants**
09h15 - 10h00	**Travaux pratiques sur bases de données de test : mesure de la l-diversité avec SPSS, Excel ou SQL** (Blum Dominique) Questions des participants
10h00 - 10h20	**Pause-café**
10h20 - 12h30	**Travaux pratiques sur bases de données de test : impact du floutage sur la mesure de la l-diversité** (Blum Dominique) **Synthèse des applications pratiques** (Blum Dominique) Questions des participants
12h30 - 14h00	**Déjeuner**
14h00 - 16h30	**Techniques de réduction du risque de ré-identification** (Blum Dominique)
16h30 - 17h00	Questions des participants **Pause-café**

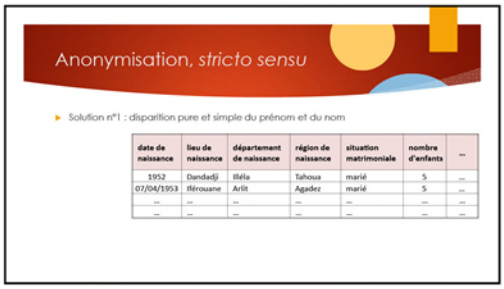
Vendredi 9 Mai 2018 – Hôtel Terminus	
09h00 - 09h15	**Installation des participants**
09h15 - 10h00	**Applications pratiques de réduction du risque de ré-identification** (Blum Dominique) Questions des participants
10h00 - 10h20	**Pause-café**
10h20 - 12h30	**Questions posées par la diffusion d'une base de microdata** (Blum Dominique) Questions des participants
12h30 - 14h00	**Déjeuner**
14h00 - 16h30	**Récapitulatif des cinq journées** (Blum Dominique) **Recommandations générales** (Blum Dominique) **Synthèse des cinq journées – Tour de table général** Questions des participants
16h30 - 17h00	**Pause-café**

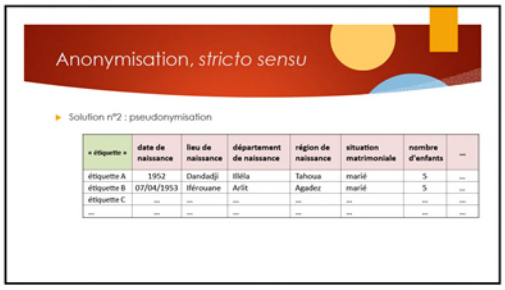
4.3 Annexe 3 : Supports de formation



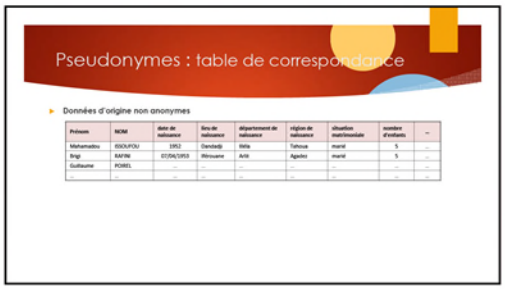


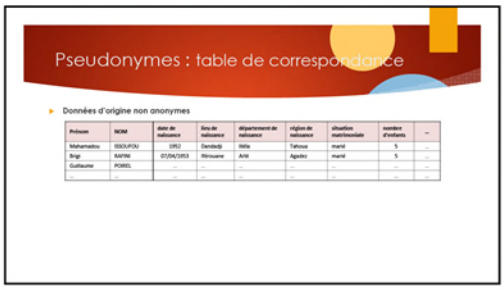


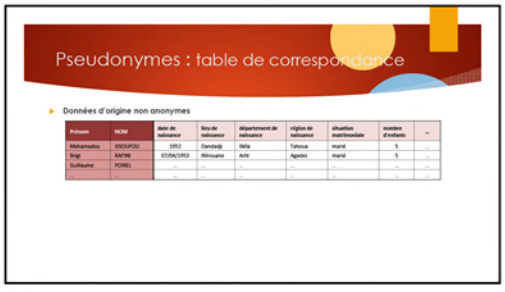


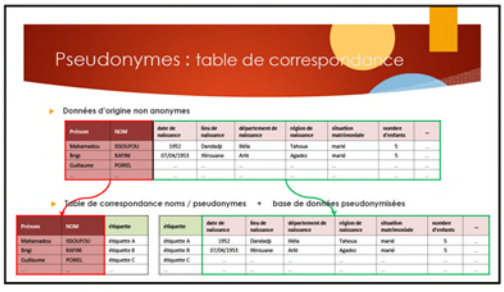

Anonymisation, *stricto sensu*
Solution n°2 : pseudonymisation
« étiquette »
date de naissance
lieu de naissance
département de naissance
région de naissance
situation matrimoniale
nombre d'enfants
étiquette A
1952
Dandadji
Illéla
Tahoua
marié
5
étiquette B
07/04/1953
Iférouane
Arlit
Agadez
marié
5
étiquette 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

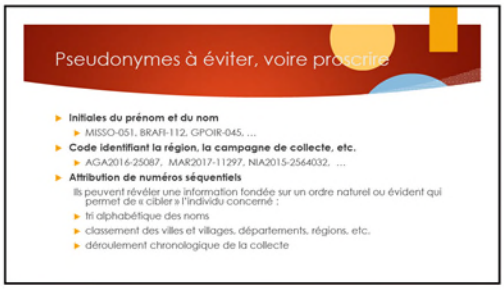

Pseudonymes : table de correspondance
Données d'origine non anonymes

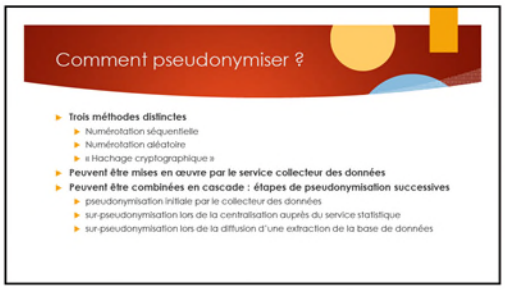

Pseudonymes : table de correspondance
Données d'origine non anonymes

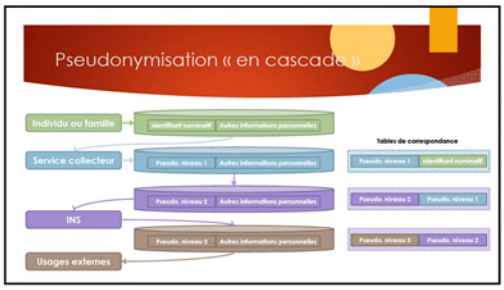

Pseudonymes : table de correspondance
Données d'origine non anonymes
Table de correspondance noms / pseudonymes + base de données pseudonymisées

2


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 : totalement 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 »
Individu ou famille
Service collecteur
INS
Usages externes
Tables de correspondance

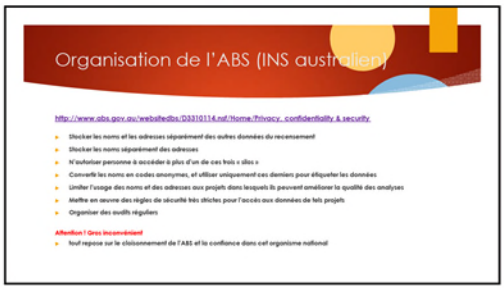

Pseudonymisation : méthode n°1
Numéro séquentiel
uniquement si les données ne sont pas fournies selon un tri préalable fondé sur un ordre naturel ou évident (ordre alphabétique des noms, classement habituel des villes et villages du département, des départements de la région, des régions du Niger, ordre chronologique de la collecte des données, etc.)
peut être mis en œuvre par le service collecteur des données


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


Pseudonymisation : méthode n°3
« 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°3
« Message en entrée » constitué des informations nominatives
exemple MAMADOUISSOUFOU
Application de la fonction de hachage sur le « message »
« Clef de hachage » en sortie
exemple 9p882kM145bnDj3b075MmAe22yfN113FdF4n63967mAe537f7z8582qeZwSQ45
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)

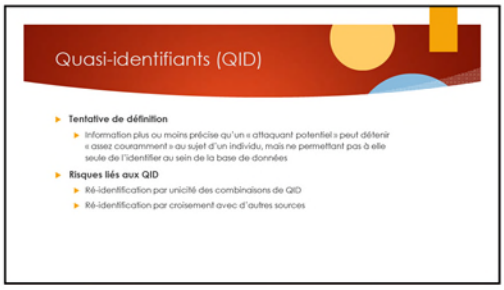

Conserver la correspondance
Pourquoi conserver la table de correspondance ?
pour « remonter » de l'enregistrement pseudonymisé au recueil d'origine
pour « redescendre » du recueil d'origine à l'enregistrement pseudonymisé
sauf dans le cas du hachage cryptographique
Qui doit conserver la table de correspondance ?
l'organisme qui est « en amont »
Comment conserver la table de correspondance ?
s'inspirer de l'organisation de l'Australian bureau of statistics (ABS)

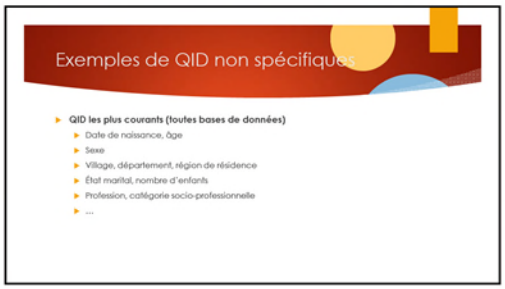

Organisation de l'ABS (INS australien)
http://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Privacy_confidentiality & security
Stocker les noms et les adresses séparément des autres données du recensement
Stocker les noms séparément des adresses
N'autoriser personne à accéder à plus d'un de ces trois « silos »
Convertir les noms en codes anonymes, et utiliser uniquement ces derniers pour étiqueter les données
Limiter l'usage des noms et des adresses aux projets dans lesquels ils peuvent améliorer la qualité des analyses
Mettre en œuvre des règles de sécurité très strictes pour l'accès aux données de tels projets
Organiser des audits réguliers
Attention ! Gros inconvénient
Tout repose sur le cloisonnement de l'ABS et la confiance dans cet organisme national


Catégories de données
vis-à-vis de l'anonymisation
•IDENTIFIANTS NOMINATIFS
•DONNÉES SENSIBLES
•QUASI-IDENTIFIANTS
•AUTRES DONNÉES


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


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

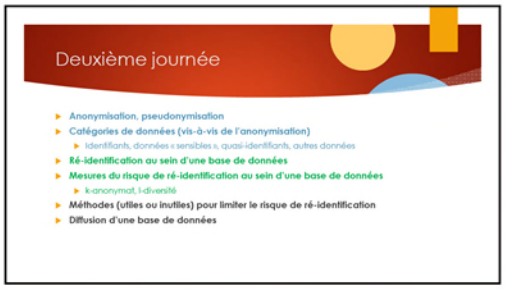

Quasi-identifiants (QID)
Tentative de définition
Information plus ou moins précise qu'un « attaquant potentiel » peut détenir « assez couramment » au sujet d'un individu, mais ne permettant pas à elle seule de l'identifier au sein de la base de données
Risques liés aux QID
Ré-identification par unicité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


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 de QID uniques


Anonymisation des données
théorie et pratique
PLATEFORMES NATIONALES D'INFORMATION POUR LA NUTRITION
NATIONAL INFORMATION PLATFORMS FOR NUTRITION
NIPN


Deuxième journée
Anonymisation, pseudonymisation
Catégories de données (vis-à-vis de l'anonymisation)
Identifiants, données « sensibles », quasi-identifiants, autres données
Ré-identification au sein d'une base de données
Mesures du risque de ré-identification au sein d'une base de données
k-anonymat, l-diversité
Méthodes (utiles ou inutiles) pour limiter le risque de ré-identification
Diffusion d'une base de données


Ré-identification
au sein d'une base de données

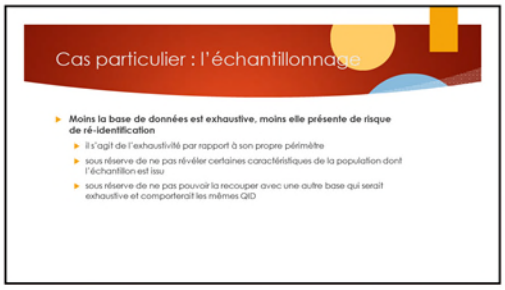

Qu'entend-on par « ré-identification » ?
Retrouver au sein d'une base de données les enregistrements relatifs à une personne dont on connaît déjà certains QID
Retrouver des combinaisons de QID dont on dispose dans d'autres bases de données, qui fournissent l'identité des personnes concernées
cf. 1997 : 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
1er cas : risque empirique faible
résoudre la question des cas extrêmes (outliers)
2e 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

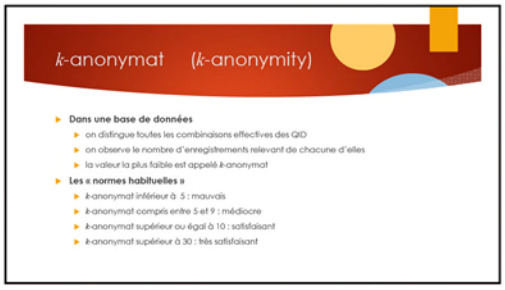
6

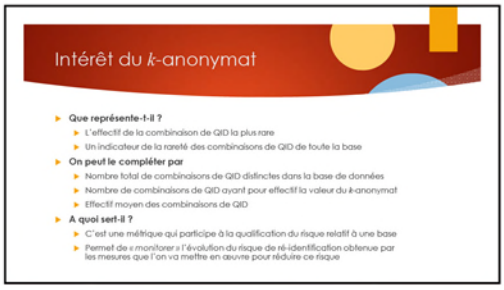

Cas particulier : l'échantillonnage
Moins la base de données est exhaustive, moins elle présente de risque de ré-identification
il s'agit de l'exhaustivité par rapport à son propre périmètre
sous réserve de ne pas révéler certaines caractéristiques de la population dont l'échantillon est issu
sous réserve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


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 dires 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

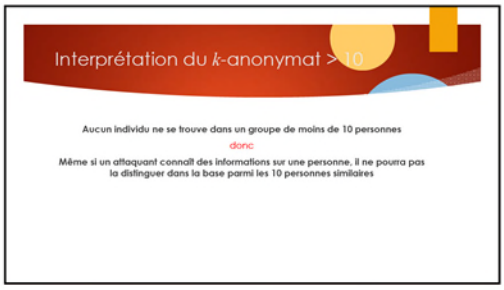

Anonymisation des données
Anonymisation, pseudonymisation
Catégories de données (vis-à-vis de l'anonymisation)
Identifiants, données « sensibles », quasi-identifiants, autres données
Ré-identification au sein d'une base de données
Mesures du risque de ré-identification au sein d'une base de données
k-anonymat, l-diversité
Méthodes (utiles ou inutiles) pour limiter le risque de ré-identification
Diffusion d'une base de données


Mesures du risque de ré-identification
au sein d'une base de données
k-ANONYMAT
l-DIVERSITÉ


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

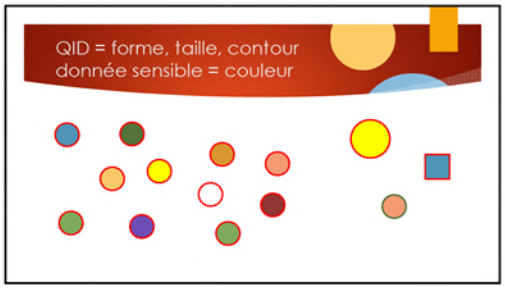

Intérêt du k-anonymat
Que représente-t-il ?
L'effectif de la combinaison de QID la plus rare
Un indicateur de la rareté des combinaisons de QID de toute la base
On peut le compléter par
Nombre total de combinaisons de QID distinctes dans la base de données
Nombre de combinaisons de QID ayant pour effectif 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

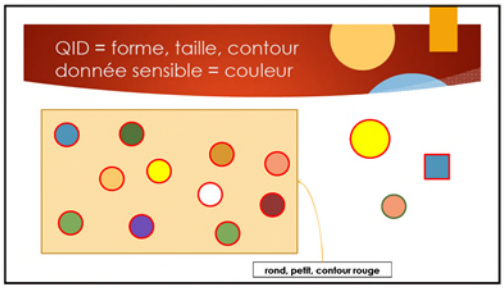

Interprétation du k-anonymat > 10
Aucun individu ne se trouve dans un groupe de moins de 10 personnes


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

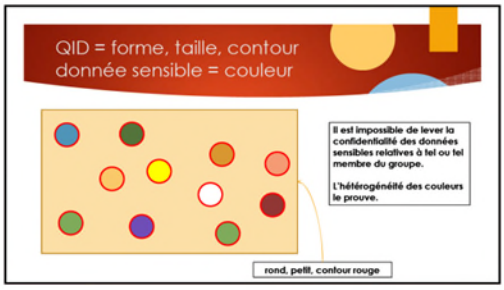

Interprétation du k-anonymat > 10
Aucun individu ne se trouve dans un groupe de moins de 10 personnes
donc
Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires
donc
Il est impossible pour un attaquant de tirer des 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

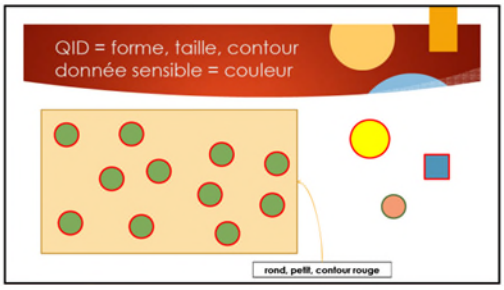

Interprétation du k-anonymat > 10
Aucun individu ne se trouve dans un groupe de moins de 10 personnes
donc
Même si un attaquant connaît des informations sur une personne, il ne pourra pas la distinguer dans la base parmi les 10 personnes similaires
donc
Il est impossible pour un attaquant de tirer des 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 !

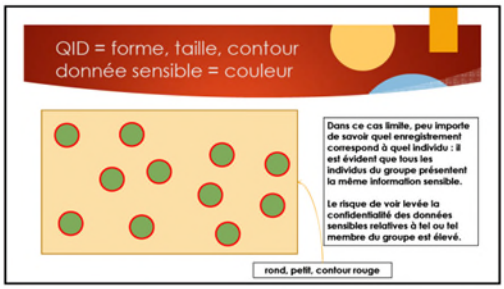

QID = forme, taille, contour
donnée sensible = couleur

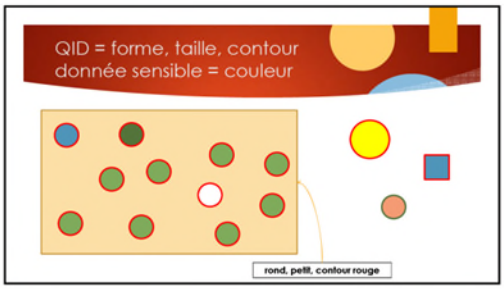

QID = forme, taille, contour
donnée sensible = couleur
rond, petit, contour rouge

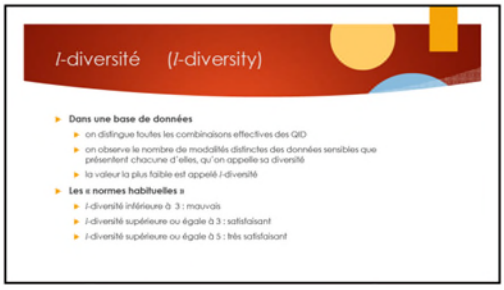
8

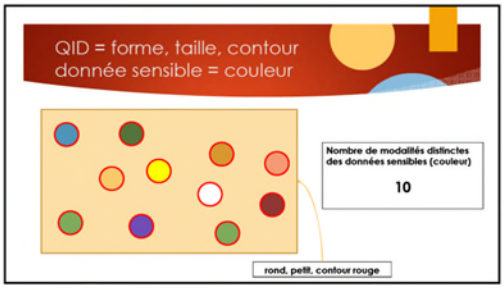

QID = forme, taille, contour
donnée sensible = couleur
Il est impossible de lever la confidentialité des données sensibles relatives à tel ou tel membre du groupe.
L'hétérogénéité des couleurs le prouve.
rond, petit, contour rouge

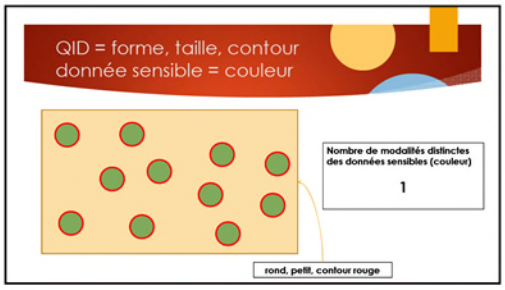

QID = forme, taille, contour
donnée sensible = couleur
rond, petit, contour rouge

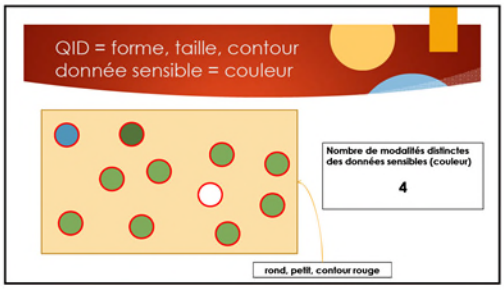

QID = forme, taille, contour
donnée sensible = couleur
Dans ce cas limite, peu importe de savoir quel enregistrement correspond à quel individu : il est évident que tous les individus du groupe présentent la même information sensible.
Le risque de voir levée la confidentialité des données sensibles relatives à tel ou tel membre du groupe est élevé.
rond, petit, contour rouge


QID = forme, taille, contour
donnée sensible = couleur
rond, petit, contour rouge

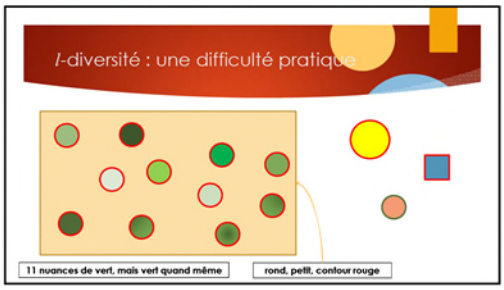

l-diversité (*l*-diversity)
Dans une base de données
on distingue toutes les combinaisons effectives des QID
on observe le nombre de modalités distinctes des données sensibles que présentent chacune d'elles, qu'on appelle sa diversité
la valeur la plus faible est appelé *l*-diversité
Les « normes habituelles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


QID = forme, taille, contour
donnée sensible = couleur
Nombre de modalités distinctes des données sensibles (couleur)
10
rond, petit, contour rouge

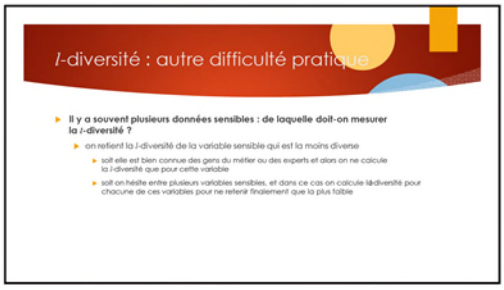

QID = forme, taille, contour
donnée sensible = couleur
Nombre de modalités distinctes des données sensibles (couleur)
1
rond, petit, contour rouge


QID = forme, taille, contour
donnée sensible = couleur
Nombre de modalités distinctes des données sensibles (couleur)
4
rond, petit, contour rouge


Intérêt de la l-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

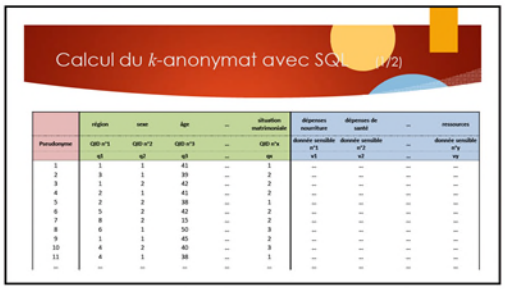

l-diversité : une difficulté pratique
11 nuances de vert, mais vert quand même
rond, petit, contour rouge

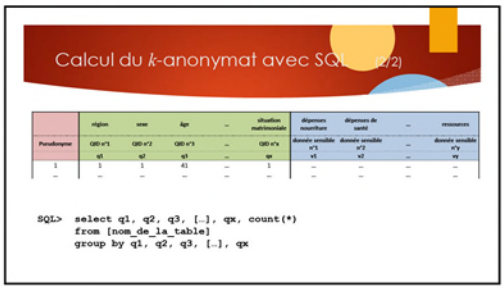

l-diversité : une difficulté pratique
Les modalités précises des données sensibles peuvent être distinctes mais très proches
un attaquant pourrait finalement tirer des conclusions « assez précises » sur tous les individus qui composent le groupe
la confidentialité des informations relatives à tous les 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)


l-diversité : autre difficulté pratique
Il y a souvent plusieurs données sensibles : de laquelle doit-on mesurer la l-diversité ?
on retient la l-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 l-diversité que pour cette variable
soit on hésite entre plusieurs variables sensibles, et dans ce cas on calcule la-diversité pour chacune de ces variables pour ne retenir finalement que la plus faible

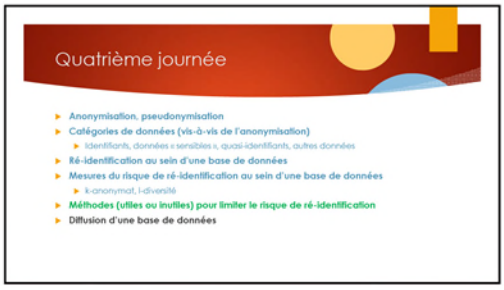

Anonymisation des données
théorie et pratique
PLATEFORMES NATIONALES D'INFORMATION POUR LA NUTRITION
NATIONAL INFORMATION PLATFORMS FOR NUTRITION
NiPN

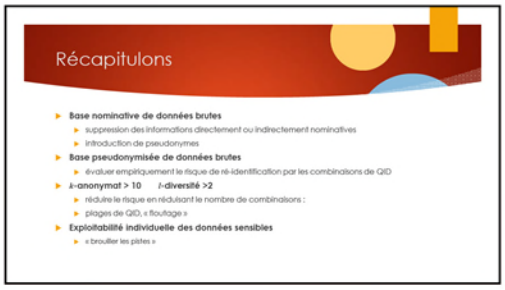

Troisième journée
Travaux pratiques


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


Calcul du k-anonymat avec SQL (2/2)
SQL> select q1, q2, q3, […], qx, count(*)
from [nom_de_la_table]
group by q1, q2, q3, […], qx


Anonymisation des données
théorie et pratique
PLATEFORMES NATIONALES D'INFORMATION POUR LA NUTRITION
NATIONAL INFORMATION PLATFORMS FOR NUTRITION
NiPN


Quatrième journée
Anonymisation, pseudonymisation
Catégories de données (vis-à-vis de l'anonymisation)
Identifiants, données « sensibles », quasi-identifiants, autres données
Ré-identification au sein d'une base de données
Mesures du risque de ré-identification au sein d'une base de données
k-anonymat, l-diversité
Méthodes (utiles ou inutiles) pour limiter le risque de ré-identification
Diffusion d'une base de données

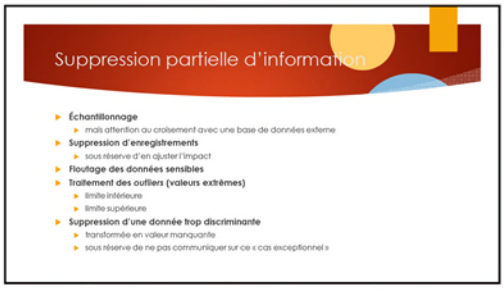

Récapitulons
Base nominative de données brutes
suppression des informations directement ou indirectement nominatives
introduction de pseudonymes
Base pseudonymisée de données brutes
évaluer empiriquement le risque de ré-identification par les combinaisons de QID
k-anonymat > 10 l-diversité >2
réduire le risque en réduisant le nombre de combinaisons :
plages de QID, « floutage »
Exploitabilité individuelle des données sensibles
« brouiller les pistes »

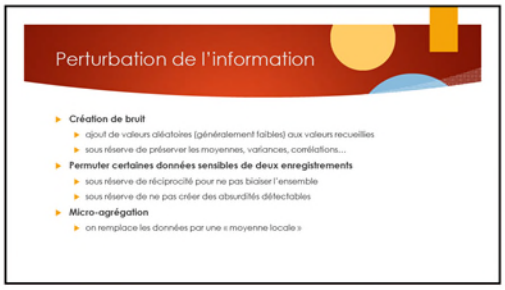

Réduction des risques
de ré-identification et d'exploitabilité individuelle
ACTIONS SUR LES QID
ACTIONS SUR LES DONNÉES SENSIBLES


Action sur les QID : « 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

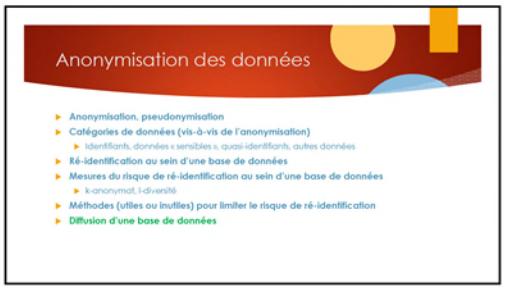

Action sur les QID : « floutage » (2/2)
Garder à l'esprit qu'en général
réduire les modalités d'un facteur de 10 réduit le nombre de groupes d'un facteur de 2, voire 3 mais pas plus
réduire les modalités des QID
agit très peu sur l'effectif de la combinaison de QID la plus rare (le k-anonymat)
mais agit 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

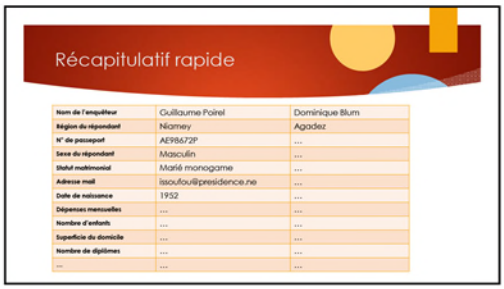

« Masquage » des données
Principe
modifier les données sensibles de certains enregistrements
tout en conservant « certaines » de leurs caractéristiques statistiques
moyennes, corrélations, etc.
Deux méthodes
suppression partielle de l'information
perturbation de l'information
Applicable aux QID et aux données sensibles

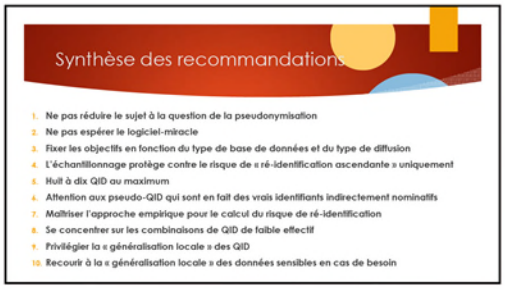

Suppression partielle d'information
Échantillonnage
mais attention au croisement avec une base de données externe
Suppression d'enregistrements
sous réserve d'en ajuster l'impact
Floutage des données sensibles
Traitement des outliers (valeurs extrêmes)
limite intérieure
limite supérieure
Suppression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

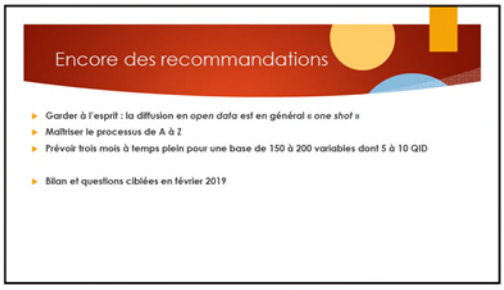

Perturbation de l'information
Création de bruit
ajout de valeurs aléatoires (généralement faibles) aux valeurs recueillies
sous réserve de préserver les moyennes, variances, corrélations…
Permuter certaines données sensibles de deux enregistrements
sous réserve de réciprocité pour ne pas biaiser l'ensemble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

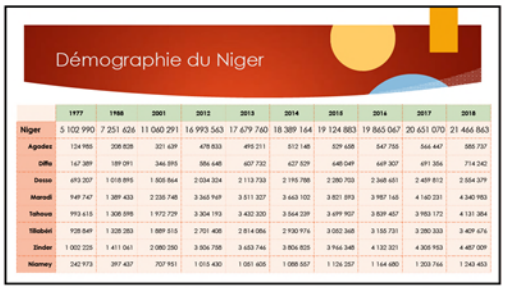

Anonymisation des données
théorie et pratique
PLATEFORMES NATIONALES D'INFORMATION POUR LA NUTRITION
NATIONAL INFORMATION PLATFORMS FOR NUTRITION
NiPN


Anonymisation des données
Anonymisation, pseudonymisation
Catégories de données (vis-à-vis de l'anonymisation)
Identifiants, données « sensibles », quasi-identifiants, autres données
Ré-identification au sein d'une base de données
Mesures du risque de ré-identification au sein d'une base de données
k-anonymat, l-diversité
Méthodes (utiles ou inutiles) pour limiter le risque de ré-identification
Diffusion d'une base de données


Récapitulatif rapide
Nom de l'enquêteur | Guillaume Poirel | Dominique Blum
Région du répondant | Niamey | Agadez
N° de passeport | AE98672P | …
Sexe du répondant | Masculin | …
Statut matrimonial | Marié monogame | …
Adresse mail | issoufou@presidence.ne | …
Date de naissance | 1952 | …
Dépenses mensuelles | … | …
Nombre d'enfants | … | …
Superficie du domicile | … | …
Nombre de diplômes | … | …
… | … | …


Synthèse des 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 protège contre le risque de « ré-identification ascendante » uniquement
5. Huit à dix QID au maximum
6. Attention aux pseudo-QID 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 QID de faible effectif
9. Privilégier la « généralisation locale » des QID
10. Recourir à la « généralisation locale » des données sensibles en cas de besoin


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 QID
Bilan et questions ciblées en février 2019



	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	124 985	208 828	321 639	478 833	495 211	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	1 506 864	2 034 324	2 113 733	2 195 786	2 280 703	2 368 651	2 459 812	2 554 379
Maradi	949 747	1 389 433	2 236 748	3 365 969	3 511 327	3 663 102	3 821 593	3 987 165	4 160 231	4 340 953
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	4 131 384
Tillabéri	928 849	1 328 283	1 889 515	2 701 408	2 814 086	2 930 976	3 052 368	3 178 731	3 310 333	3 409 676
Zinder	1 002 226	1 411 061	2 080 250	3 306 788	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