



PLATEFORME
DES DONNÉES
DE SANTÉ
FRENCH HEALTH DATA HUB

Plateforme des données de santé

Guide pédagogique

Appariements à la base
principale du SNDS



Objectifs



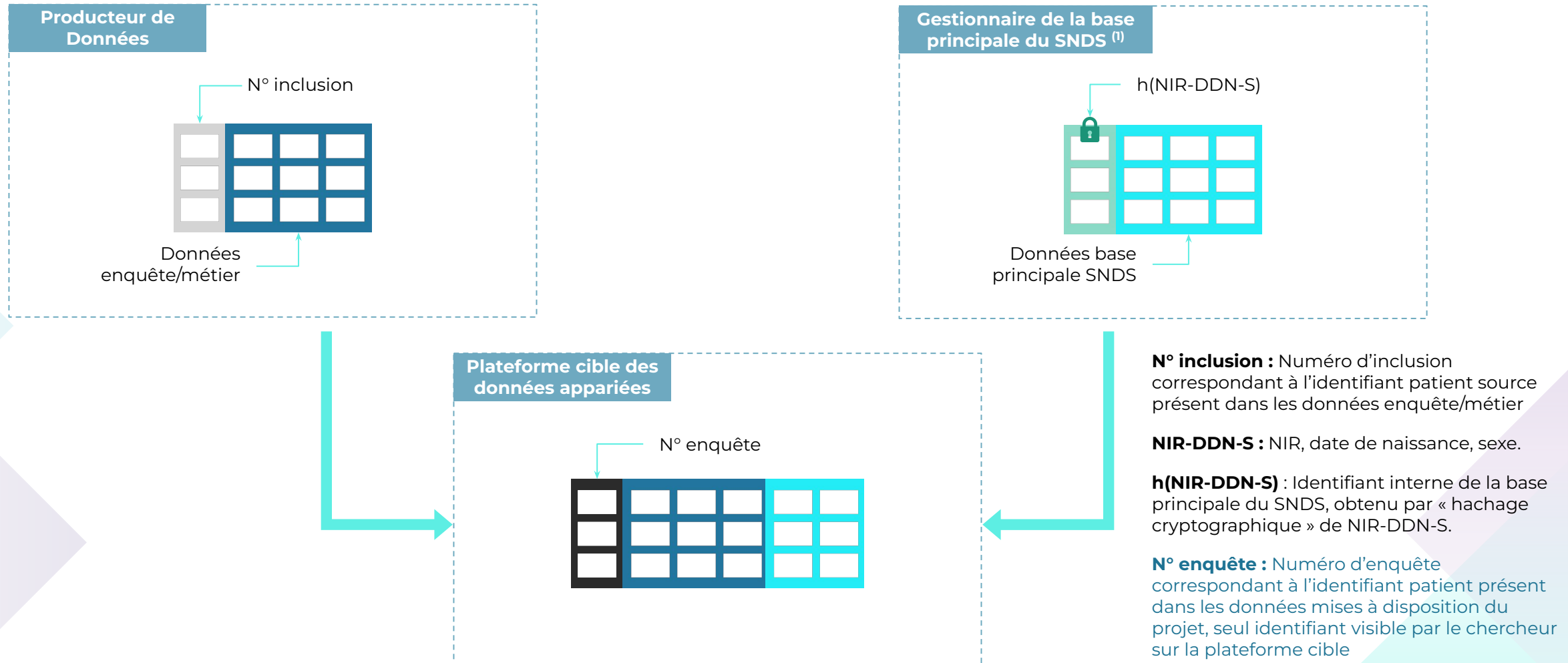
L'objectif de ce document est d'aider les porteurs de projet et producteurs de données à appréhender les différentes possibilités d'appariement à la base principale du Système National des Données de Santé (SNDS), et à choisir la modalité d'appariement pertinente dans leur situation, le cas échéant.

Pour ce faire, ce guide contient notamment une présentation synthétique des différents types d'appariement, avec pour chacun :

- Le circuit suivi par les données
- Les principaux avantages
- Les principales étapes de préparation
- Des ressources documentaires pour aller plus loin

Un appariement à la base principale du SNDS

Rapprochement d'ensembles de données distincts à l'aide d'informations communes



La base principale du SNDS et l'identifiant bénéficiaire

Comment est construit l'identifiant bénéficiaire dans le SNDS ?

La base principale du SNDS mobilise, pour un même bénéficiaire, deux systèmes différents d'identifiants :

- Le **pseudo-NIR**, créé pour la base principale du SNDS et utilisé dans les tables sous la forme pseudonymisée $h(\text{NIR-DDN-S})$. Il s'agit de l'**identifiant interne de la base principale du SNDS**. Un individu change généralement au cours de sa vie de pseudo-NIR, car celui-ci est composé notamment du NIR de l'ouvreur de droit auquel le bénéficiaire est rattaché.
- Le **NIR du bénéficiaire** (numéro de sécurité sociale habituel), qui n'est utilisé dans la base principale du SNDS (sous forme pseudonymisée) que pour relier différents pseudo-NIR d'un même bénéficiaire. Cet identifiant est pérenne et unique au bénéficiaire, mais n'est pas toujours renseigné dans la base principale du SNDS (environ 5% de cas manquants).

Il est donc important, lorsque l'on cherche à reconstituer le parcours de soin d'un bénéficiaire dans la base principale du SNDS, de mobiliser conjointement le(s) pseudo-NIR et le NIR du bénéficiaire, qui ensemble forment les **Composantes à Pseudonymiser**.

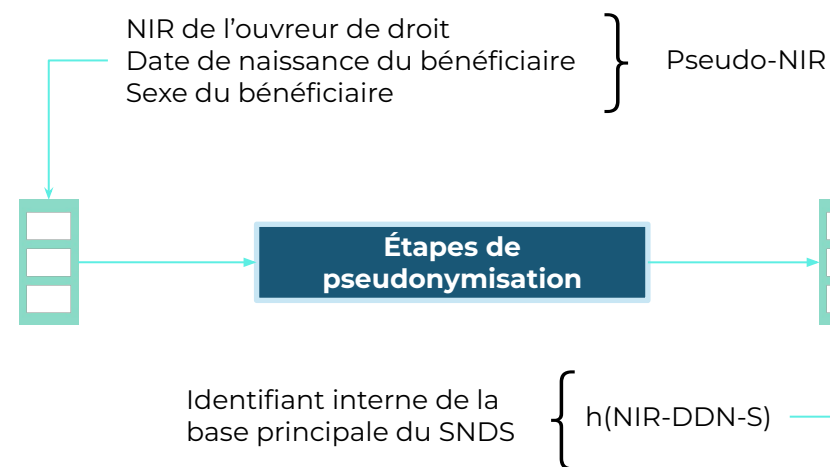
La pseudonymisation des identifiants

Afin de protéger la confidentialité des données de la base principale du SNDS, les identifiants des bénéficiaires sont pseudonymisés avant d'être associés aux données : ainsi, nul ne peut disposer simultanément du NIR en clair et des données associées à ce NIR sur l'ensemble de la base principale du SNDS.

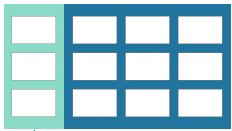
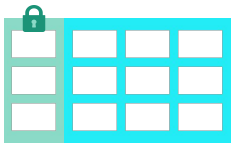
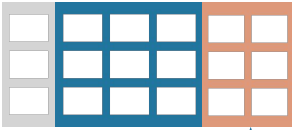
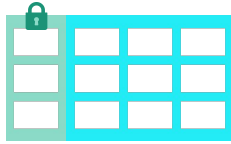
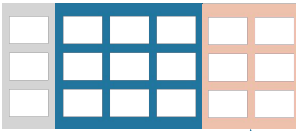
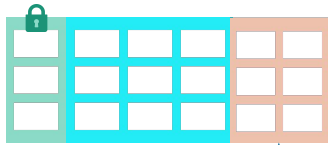
Pour ce faire, une fonction de hachage à clef secrète est appliquée aux Composantes à Pseudonymiser, permettant de pseudonymiser de façon irréversible ces identifiants. Pour garantir un niveau plus élevé de protection des données, cet algorithme de pseudonymisation est appliqué lors de plusieurs étapes successives, par des acteurs différents utilisant des secrets distincts. Ce circuit de pseudonymisation dans son ensemble est orchestré par la CNAM.

Pour en savoir plus

- Fiche "[Identifiant des bénéficiaires](#)" dans la documentation collaborative



Les différentes possibilités d'appariement

Possibilité	Situation	Étapes majeures	Commentaire
Appariement direct (déterministe) par l'utilisation du NIR	 	<ol style="list-style-type: none"> Extraction des données enquête/métier Pseudonymisation du NIR Extraction base principale SNDS Mise à disposition des données appariées 	<p>Modalités de transfert des données Canal sécurisé (données enquête/métier) Solution "Mon interface appariement" (NIR)</p> <p>Acteurs supplémentaires ⁽¹⁾ PDS, CNAM</p> <p>Caractéristiques Très bonne qualité d'appariement Simplicité et délais réduits Nécessite le NIR</p>
Appariement direct (déterministe) avec reconstitution du NIR	 	<ol style="list-style-type: none"> Extraction des données enquête/métier Reconstitution du NIR à partir de l'état civil Pseudonymisation du NIR Extraction base principale SNDS Mise à disposition des données appariées 	<p>Modalités de transfert des données Canal sécurisé (données enquête/métier) Solution "Mon interface appariement" (données état civil)</p> <p>Acteurs supplémentaires ⁽¹⁾ PDS, CNAM, CNAV (SNGI)</p> <p>Caractéristiques Bonne qualité d'appariement Simplicité et délais réduits Nécessite l'état civil</p>
Appariement indirect (probabiliste) sur des variables communes	 	<ol style="list-style-type: none"> Extraction des données enquête/métier Croisement des variables communes Extraction base principale SNDS Mise à disposition des données appariées 	<p>Modalités de transfert des données Canal sécurisé (données enquête/métier) BlueFiles - CNAM (variables communes)</p> <p>Acteurs supplémentaires ⁽¹⁾ PDS, CNAM</p> <p>Caractéristiques Flexibilité Qualité d'appariement variable</p>

⁽¹⁾ Dans chaque situation, le gestionnaire du SNDS et l'hébergeur du projet sont nécessairement impliqués



Appariement direct (déterministe) par l'utilisation du NIR

Préparer l'appariement direct par l'utilisation du NIR

Dans quel cas choisir l'appariement direct par l'utilisation du NIR?

L'appariement direct (déterministe) par l'utilisation du NIR est la méthode offrant *a priori* le meilleur taux d'appariement. En conséquence, ce type d'appariement doit être privilégié dès lors que la base de données enquête/métier source contient, de manière fiable, le NIR du bénéficiaire et les composantes du pseudo-NIR (NIR de l'ouvreur de droit, date de naissance du bénéficiaire, sexe du bénéficiaire) permettant de reconstruire l'identifiant interne de la base principale du SNDS h(NIR-DDN-S).

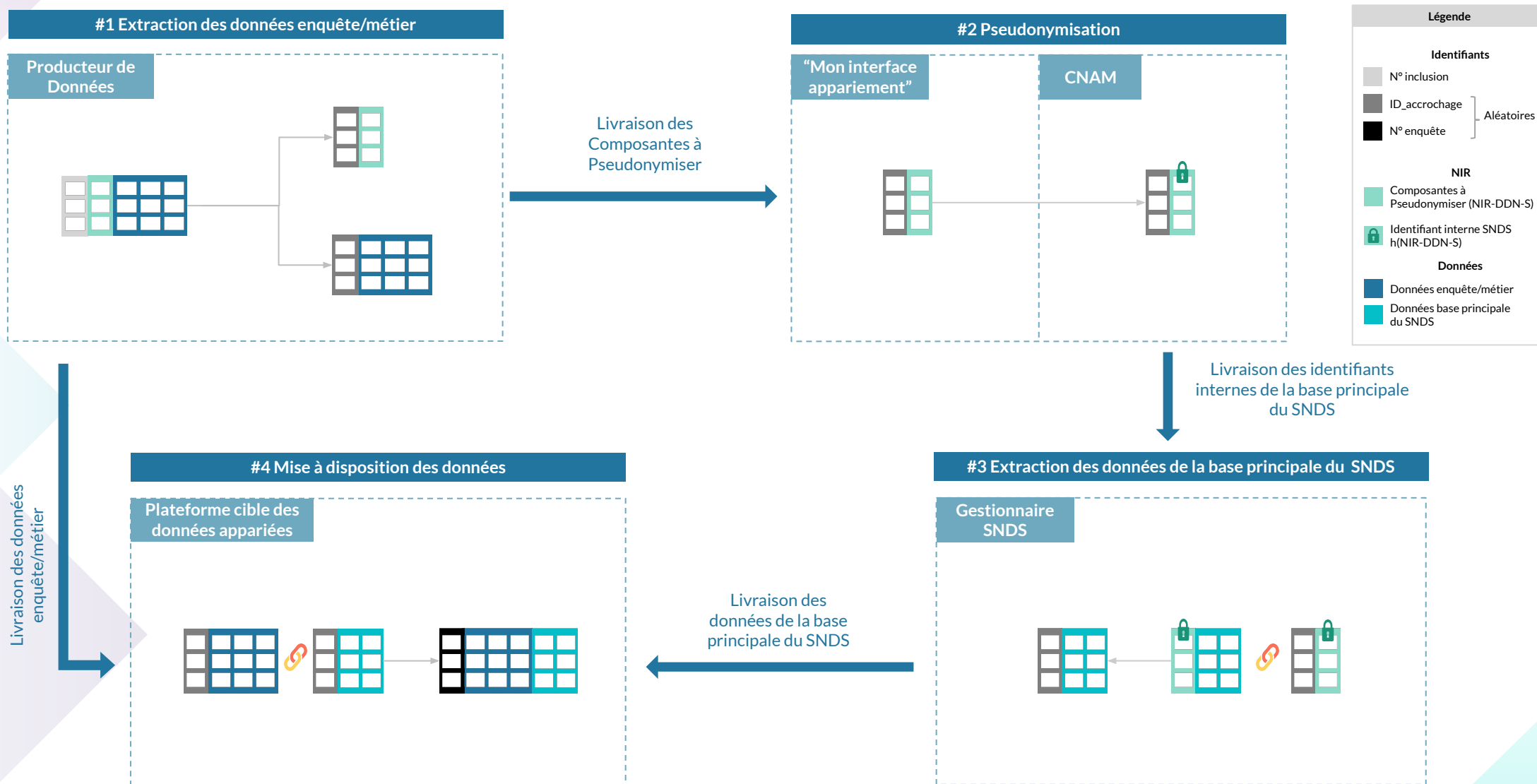
Étapes de préparation

Constitution	Nettoyage	Dépôt sur "Mon interface appariement"
<p>La table d'accrochage est une extraction de la base des données enquête/métier source présentant les propriétés suivantes :</p> <ul style="list-style-type: none">• Elle contient pour chaque individu les Composantes à Pseudonymiser, i.e. :<ul style="list-style-type: none">○ NIR de l'ouvreur de droit○ Date de naissance du bénéficiaire○ Sexe du bénéficiaire○ NIR du bénéficiaire• Elle contient un identifiant non signifiant par individu, créé spécifiquement pour le projet et différent de l'identifiant présent dans la base source (ID_acc).	<p>La table d'accrochage doit être nettoyée avec soin, car de la qualité de la table dépend directement le taux d'appariement finalement atteint.</p> <p>Cette phase de nettoyage implique typiquement de vérifier la complétude, l'exactitude et la plausibilité des données.</p>	<p>Une fois constituée et nettoyée, et avant d'être déposée sur "Mon interface appariement", la table d'accrochage doit être convertie en format csv, puis chiffrée avec la clé de chiffrement de "Mon interface appariement" et signée avec la clé de signature du Producteur de données qui fait la demande.</p> <p>Pour ce faire, la Plateforme des Données de Santé met à disposition une documentation sur l'interface utilisateur du "Mon interface appariement".</p>

Pour en savoir plus

- La CNIL a publié en 2020 un guide pratique intitulé [Modalités de circulation du NIR pour la recherche en santé aux fins d'appariement de données avec le SNDS](#) couvrant les variantes principales de ce circuit.
- Ce guide a été complété en 2024 par des [fiches pratiques](#) détaillant les différents circuits possibles.

Circuit





Appariement direct (déterministe) avec reconstitution du NIR

Préparer l'appariement direct avec reconstitution du NIR

Dans quel cas choisir l'appariement direct avec reconstitution du NIR ?

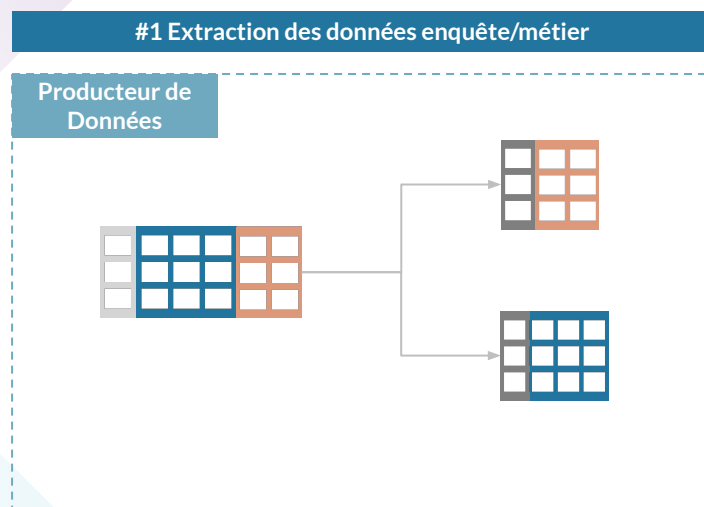
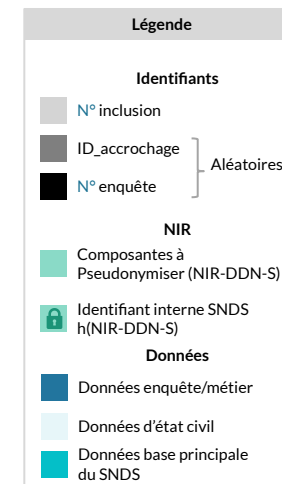
Le circuit d'appariement direct (déterministe) avec reconstitution du NIR a pour objectif de mobiliser les informations sur l'état civil des bénéficiaires (nom, prénom, sexe, date de naissance) pour retrouver leur NIR, et de se ramener ainsi au cas de l'appariement direct. Il existe en effet des systèmes permettant d'effectuer une telle opération, notamment le Système National de Gestion des Identifiants (SNGI) géré par la CNAV.

Sous réserve de disposer des informations d'état civil suffisantes, l'appariement avec reconstitution du NIR offre *a priori* un taux d'appariement comparable à l'appariement direct sur le NIR. En conséquence, ce type d'appariement doit être privilégié dès lors que deux conditions sont réunies : (1) la base source ne contient pas (ou pas de manière suffisamment fiable) les composantes du pseudo-NIR, et (2) la base source contient, de manière fiable, les composantes de l'état civil (nom, prénom, sexe, date de naissance) permettant de retrouver le pseudo-NIR le SNGI.

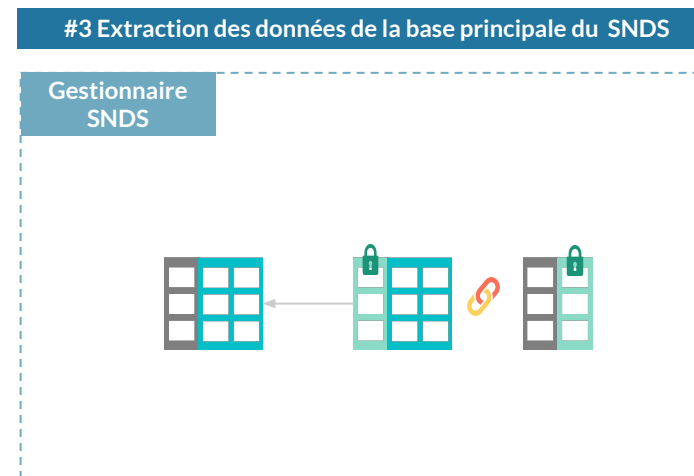
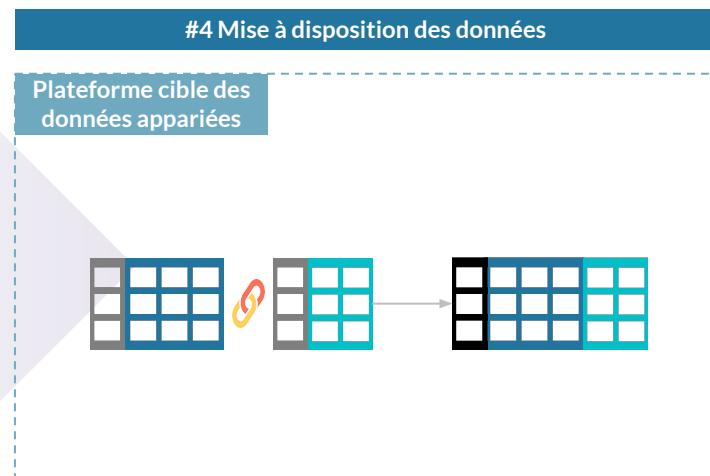
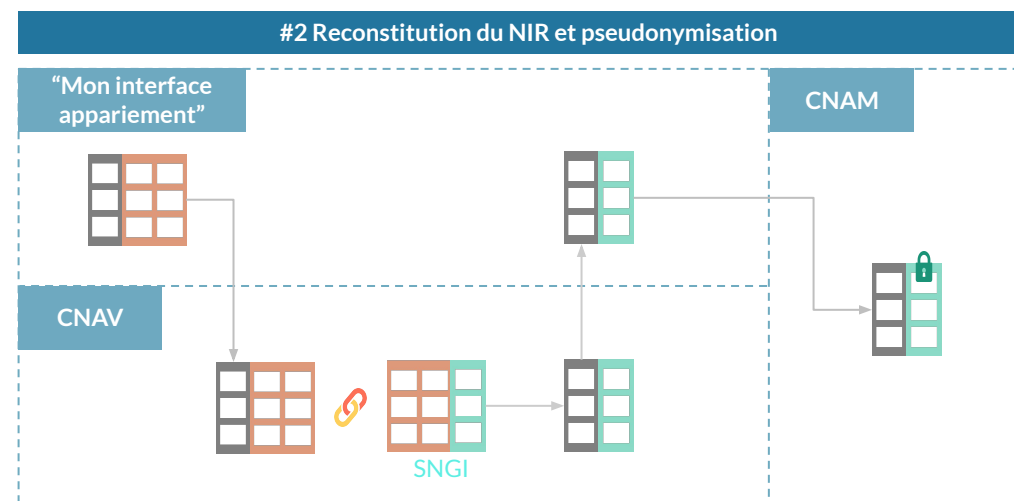
Étapes de préparation

Constitution	Nettoyage	Dépôt sur "Mon interface appariement"
<p>La table d'accrochage est une extraction de la base des données enquête/métier source présentant les propriétés suivantes :</p> <ul style="list-style-type: none">• Elle contient pour chaque individu les données d'état civil, i.e. :<ul style="list-style-type: none">◦ Nom◦ Prénom◦ Date de naissance◦ Sexe• Elle contient un identifiant non signifiant par individu, créé spécifiquement pour le projet et différent de l'identifiant présent dans la base source (ID_acc).	<p>La table d'accrochage doit être nettoyée avec soin, car de la qualité de la table dépend directement le taux d'appariement finalement atteint.</p> <p>Cette phase de nettoyage implique typiquement de vérifier la complétude, l'exactitude et la plausibilité des données..</p>	<p>Une fois constituée et nettoyée, et avant d'être déposée sur "Mon interface appariement", la table d'accrochage doit être convertie en format csv, puis chiffrée avec la clé de chiffrement de "Mon interface appariement" et signée avec la clé de signature du Producteur de données qui fait la demande.</p> <p>Pour ce faire, la Plateforme des Données de Santé met une documentation sur l'interface utilisateur du "Mon interface appariement".</p>

Circuit



Livraison des données d'état civil



Livraison des données de la base principale du SNDS

Livraison des identifiants internes de la base principale du SNDS

Livraison des données enquête/métier



Appariement indirect (probabiliste)

Préparer l'appariement indirect

Dans quel cas choisir l'appariement indirect ?

L'appariement indirect (probabiliste) offre un maximum de flexibilité, puisqu'il ne requiert pas la présence dans la base de données source de variables spécifiques qui seraient nécessaires pour reconstruire l'identifiant interne de la base principale du SNDS h(NIR-DDN-S). En contrepartie, cette méthode offre des taux d'appariement très variables en fonction de la base source des données d'enquête/métier et de la fiabilité des variables d'appariement, et un travail important doit être réalisé sur celles-ci afin d'optimiser le taux d'appariement, qui reste très difficile à estimer a priori. En conséquence, l'appariement indirect ne s'impose que lorsque deux conditions sont réunies : (1) il est impossible de procéder à un appariement direct fiable, et (2) des variables d'appariement fiables et discriminantes sont clairement identifiées dans la base source.

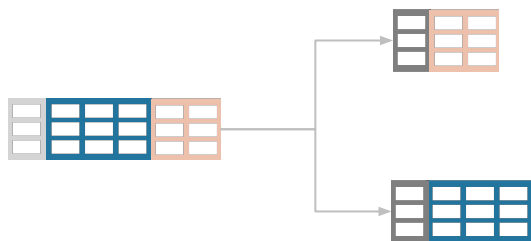
Étapes de préparation

Constitution	Nettoyage	Documentation
<p>La table d'accrochage est une extraction, idéalement au format CSV, de la base des données enquête/métier source présentant les propriétés suivantes :</p> <ul style="list-style-type: none">• Elle ne contient aucune information directement identifiante (noms, prénoms, initiales, etc...)• Elle contient un identifiant non signifiant par patient, créé spécifiquement pour le projet et différent de l'identifiant présent dans la base source (ID_acc).• Elle contient uniquement les variables nécessaires à l'appariement	<p>Cette phase de nettoyage doit être réalisée avec le plus grand soin, car de la qualité de la table dépend directement le taux d'appariement finalement atteint. Elle implique typiquement :</p> <ul style="list-style-type: none">• De corriger les valeurs aberrantes ou manquantes• Dans la mesure du possible, d'aligner les données sur les formats standards, nomenclatures et règles de codage en usage dans la base principale du SNDS. Ces nomenclatures sont disponibles dans le dictionnaire des variables.	<p>La table d'accrochage, une fois constituée et nettoyée, doit finalement être documentée de manière détaillée. Cette documentation permettra à la structure réalisant l'appariement de faire des choix cohérents dans la comparaison des données, et à chacun des acteurs impliqués de s'assurer d'une bonne compréhension du processus en cours, afin d'anticiper d'éventuelles difficultés. Pour vous assurer de documenter correctement les variables d'appariement, vous pouvez vous appuyer sur l'annexe 2 du guide EDB du Starter-Kit, qui propose une liste d'informations à vérifier et répertorier au niveau de chaque variable.</p>

Circuit

#1 Extraction des données enquête/métier

Producteur de Données



Livraison des variables communes (via BlueFiles)

#2 Extraction des données de la base principale du SNDS

Gestionnaire SNDS



Livraison des données enquête/métier

#3 Mise à disposition des données

Plateforme cible des données appariées



Livraison des données de la base principale du SNDS

Légende

Identifiants

- N° inclusion
 - ID_accrochage
 - N° enquête
- } Aléatoires

NIR

- Identifiant interne SNDS h(NIR-DDN-S)

Données

- Données enquête/métier
- Variables communes
- Données base principale du SNDS

Choisir les variables d'appariement (1/2)

Il est impossible, sans connaissance fine de la base des données enquête/métier source, de déterminer les variables qui seront les plus pertinentes, c'est pourquoi cette étape requiert un travail de fond prenant en compte simultanément les caractéristiques de la base source et celles de la base principale du SNDS. Les principes généraux qui se trouvent dans ce guide donnent un faisceau d'indices pour vous assister dans cette tâche, mais ne sauraient remplacer une réelle expertise des deux bases de données.

Principes généraux de choix

1. Comparabilité : Des informations identiques, ou directement comparables, sont encodées de part et d'autre
2. Exploitabilité : Ces informations sont exprimées dans des variables clairement identifiées et suffisamment fiables dans chaque base
3. Discrimination : Ces informations, prises ensemble, sont suffisamment discriminantes pour isoler les personnes de la population cible

Étapes et outils

1. Lister les variables exploitables dans la base de données enquête/métier source
2. Déterminer, parmi les variables exploitables, lesquelles disposent d'un équivalent dans la base principale du SNDS, sur une ou plusieurs variables (vous pouvez vous appuyer sur des outils existants tels que le [dictionnaire des variables](#) et la [documentation collaborative](#) du SNDS)
3. Répartir les variables ainsi identifiées en trois listes distinctes :
 - a. **La liste principale** (variables les plus discriminantes et fiables)
 - b. **La liste complémentaire** (variables un peu moins discriminantes ou fiables, utilisée pour améliorer le taux d'appariement)
 - c. **La liste de réserve** (variables pressenties comme les moins discriminantes, moins fiables, avec une équivalence moins claire dans la base principale du SNDS, utilisée seulement en cas de nécessité)

Visualisation de la structure du SNDS

Explorateur des variables Recherche dans les nomenclatures Explorateur des tables

? Aide 9 clés de jointure Partager la vue

Afficher 50 résultats Rechercher

Table	Variable	Libelle	Nomenclature
AI	All	All	All
BE_IDE_R	IDE_ETA_NUM	Numéro Finess de l'Etablissement	-
BE_IDE_R	IDE_ETA_NU8	Numéro Finess de l'Etabt sans clé	-
BE_IDE_R	IDE_ETA_NOM	Raison Sociale Abrégée	-
BE_IDE_R	IDE_IDE_CPL	Complément d'identification	-
BE_IDE_R	IDE_VOI_NUM	Numéro dans la voie	-
BE_IDE_R	IDE_VOI_CNU	Complément Numéro de voie	-
BE_IDE_R	IDE_VOI_TYP_LRG	Nature de la voie	-

Aperçu du dictionnaire interactif des variables de la base principale du SNDS

Choisir les variables d'appariement (2/2)

Informations typiquement utilisées

Personne

- Année et mois de naissance
- Sexe
- Date de décès
- Lieu de résidence (code commune INSEE) [attention : cette variable n'est pas totalement fiable dans le SNDS]
- Régime AM

Soins de ville

- Prestations affinées (consultations, dispositifs médicaux, actes, ...)
- Date
- Spécialité (prescripteur / exécutant)

Traitement pharmaceutique (uniquement pour les soins de ville)

- Date de prescription
- Code CIP ou ATC du traitement

Hospitalisation

- Date de début, date de fin
- FINESS de l'hôpital
- Codes CIM-10 des diagnostics
- Codes CCAM d'actes effectués
- Codes LPP de dispositifs médicaux utilisés
- Codes NABM d'examens biologiques réalisés
- GHM / GHS
- Médicaments onéreux (hors T2A)
 - Code UCD
 - Date d'administration

Dispositifs spécifiques

- ALD
- Accidents du travail, maladies professionnelles
- CMU-C

Note : Les informations surlignées sont à la fois particulièrement discriminantes, et fréquemment disponibles, et donc à considérer en priorité. « Avec pour seules informations l'hôpital, le code géographique du domicile, l'âge, le sexe, le mois de sortie et la durée du séjour, 89 % des personnes hospitalisées dans l'année 2008 (et 100 % des personnes hospitalisées deux fois cette année) sont identifiables » (Blum & Trouessin, 2012)

Note : Pour votre projet, vous pouvez vous appuyer sur la [checklist](#) pour identifier de premières correspondances avec ces variables majeures.

Informations typiquement exclues (absentes de la base principale du SNDS)

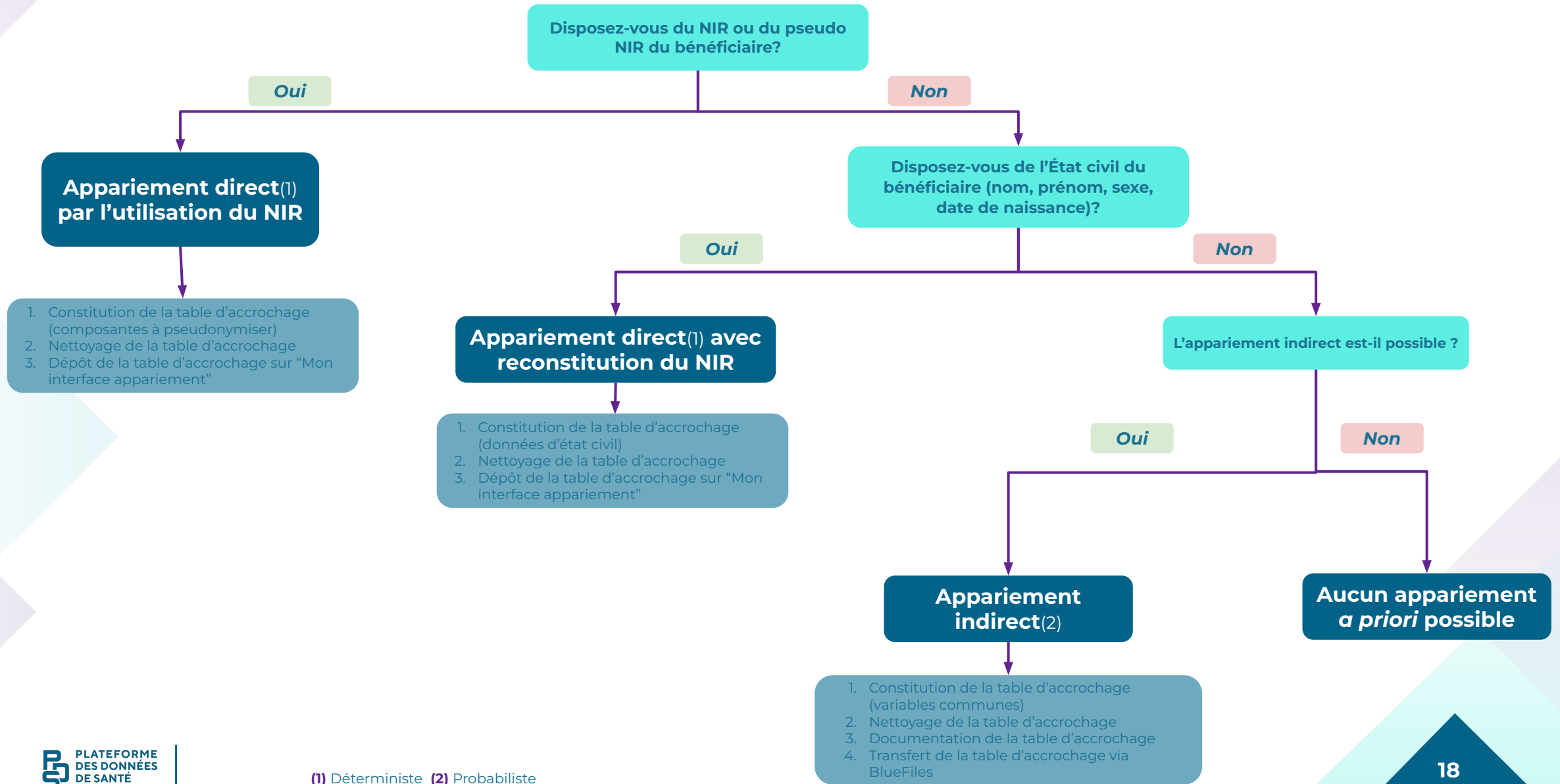
- Traitements pharmaceutiques à l'hôpital, sauf les plus onéreux
- Jour de naissance
- Résultats d'examens cliniques ou paracliniques

- Mode de vie et facteurs de risque (tabac, alcool, nutrition, ...)
- Symptômes
- Motifs de consultation ou d'interaction avec le système de santé



Choisir son type d'appariement

Synthèse du processus pour le producteur de données





Suivez-nous sur les réseaux sociaux !

