Fiche pratique

CIRCULATION DU NIR AUX FINS D'APPARIEMENT AVEC LE SNDS

CIRCUIT MULTI-CENTRES / BASE CENTRALE PSEUDONYMISÉE

Produit en collaboration avec le CASD

MC - Base pseudo – Version 1.0 – Juin 2024

CNIL.
COMMISSION NATIONALE
INFORMATIQUE & LIBERTÉS

1. Introduction

La CNIL publie un ensemble de fiches pratiques, produites avec le CASD, présentant des exemples de circuits d'appariement avec le SNDS, en complément du guide pratique de la CNIL[1] publié en décembre 2020.

Les fiches présentent :

- des schémas fonctionnels détaillés pour chaque étape, dans le même formalisme que ceux du guide ;
- des schémas techniques orientés « tables de données », produits par le CASD.

Ces fiches illustrent en détail des exemples d'implémentation concrète des circuits du guide de 2020, lequel reste valide et se trouve ainsi précisé par les fiches.

Ces exemples respectent les principes issus du guide, qui ont été déclinés pour les fiches pratiques. Vous les trouverez rassemblés dans le document vadémécum[2], comme aide-mémoire et guide de lecture.

La présente fiche concerne une étude multicentrique où les données des centres, y compris le NIR, sont stockées localement et sont en parallèle regroupées dans une base centrale pseudonymisée gérée par le RT.

À noter : contrairement à un eCRF qui dépend d'une étude spécifique, la base centrale décrite ici est constituée de manière pérenne afin de regrouper les données de différents centres en les rendant accessibles au RT.

Présentation du CASD

Le Centre d'accès sécurisé aux données (CASD) est un groupement d'intérêt public (GIP) rassemblant l'État représenté par INSEE, le GENES, le CNRS, l'École polytechnique, HEC Paris et la Banque de France.

Il a été créé par arrêté interministériel du 29 décembre 2018.

Le GIP a pour objet principal d'organiser et de mettre en œuvre des services d'accès sécurisé pour les données confidentielles à des fins non lucratives de recherche, d'étude, d'évaluation ou d'innovation. Il a également pour mission de valoriser la technologie développée pour sécuriser l'accès aux données dans le secteur public et dans le secteur privé.

[1] *Guide pratique : Modalités de circulation du NIR pour la recherche en santé aux fins d'appariement avec le SNDS* (PDF, 660 ko), CNIL, URL : https://www.cnil.fr/sites/cnil/files/atoms/files/guide_pratique_circuits_nir_recherche_en_sante.pdf

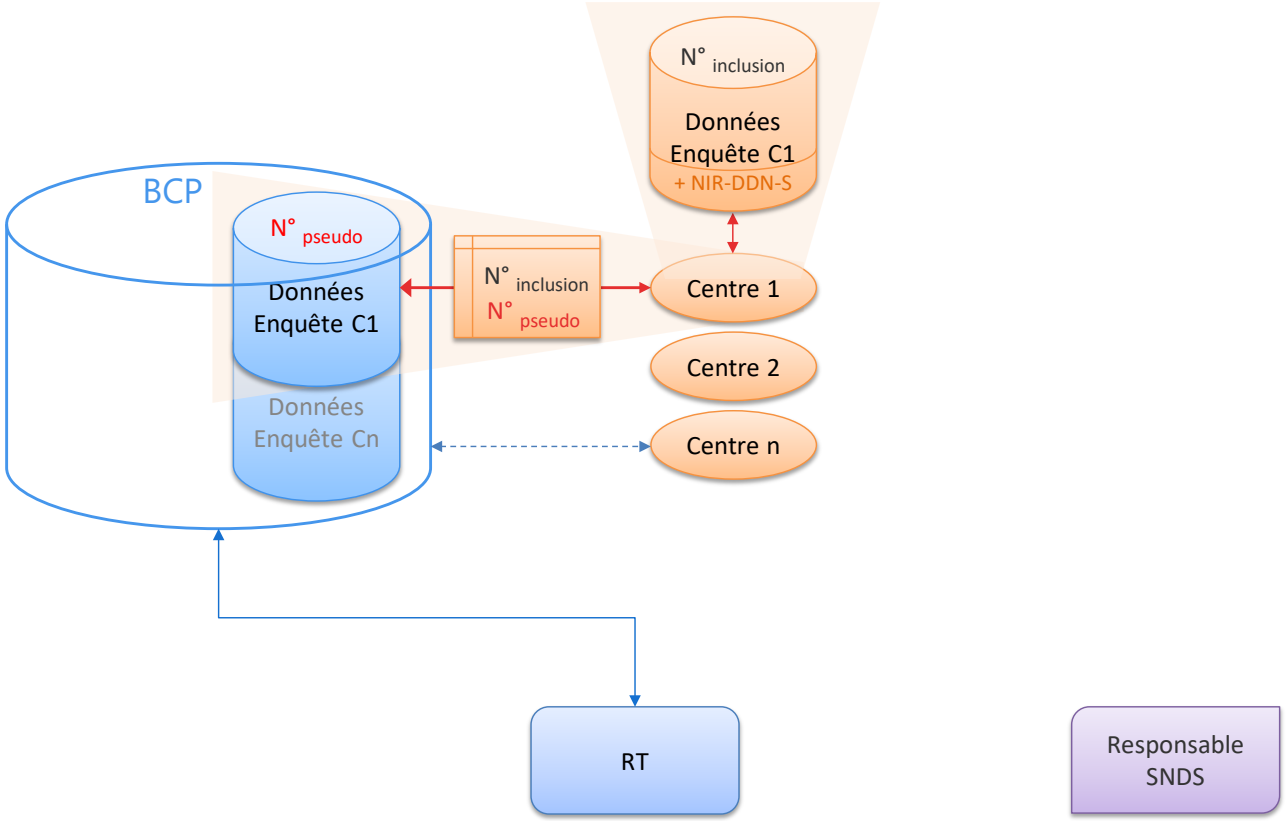
[2] *Vadémécum : circulation du NIR aux fins d'appariement avec le SNDS* (PDF, 328 ko), CNIL, URL : https://www.cnil.fr/sites/cnil/files/2024-06/circuits_nir_vademecum.pdf

2. Implémentation détaillée du circuit Multi-centres / Base centrale pseudonymisée

Centralisation des données d'enquête

Chaque centre transfère ses données locales vers une base centrale pseudonymisée gérée par le responsable de traitement

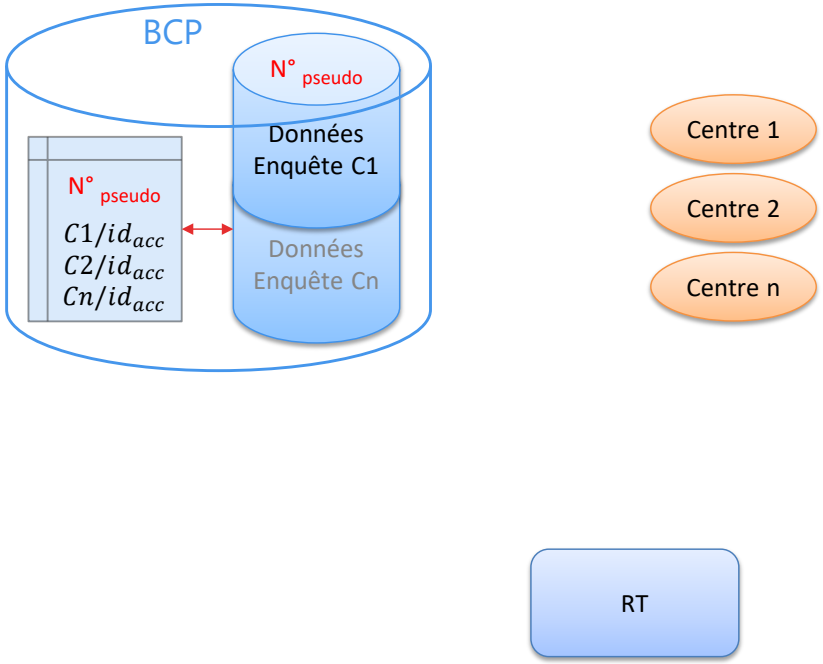
- Le responsable de traitement accède à l'ensemble des données de la base centrale pseudonymisée. Selon les cas, les centres peuvent également y consulter les données pseudonymisées de l'ensemble des centres. La sécurité est assurée par la base centrale selon les règles fixées par le RT.
- Afin de limiter les risques de réidentification, **les données des centres sont pseudonymisées avant leur insertion dans la base centrale.**
- **La table de correspondance** entre le numéro d'inclusion d'un centre et le pseudonyme de la base reste stockée dans le centre, de manière **sécurisée et cloisonnée** (par ex. dans une table séparée et chiffrée avec une clé spécifique).
- De même, **le NIR des patients est cloisonné par rapport aux données d'enquête et au numéro d'inclusion** (par exemple, le NIR est chiffré avec une clé spécifique).
- En complément de ces mesures, **une authentification forte est recommandée pour tout accès à la base centrale et à la base locale**, que ce soit en saisie de données ou pour une simple consultation.



Etape 0 – Génération des identifiants d'accrochage

La base centrale génère en interne, non visible des centres, une table de correspondance entre le numéro d'inclusion des participants et un identifiant d'accrochage aléatoire et non signifiant (C1/id_{acc} pour le premier centre, etc.)

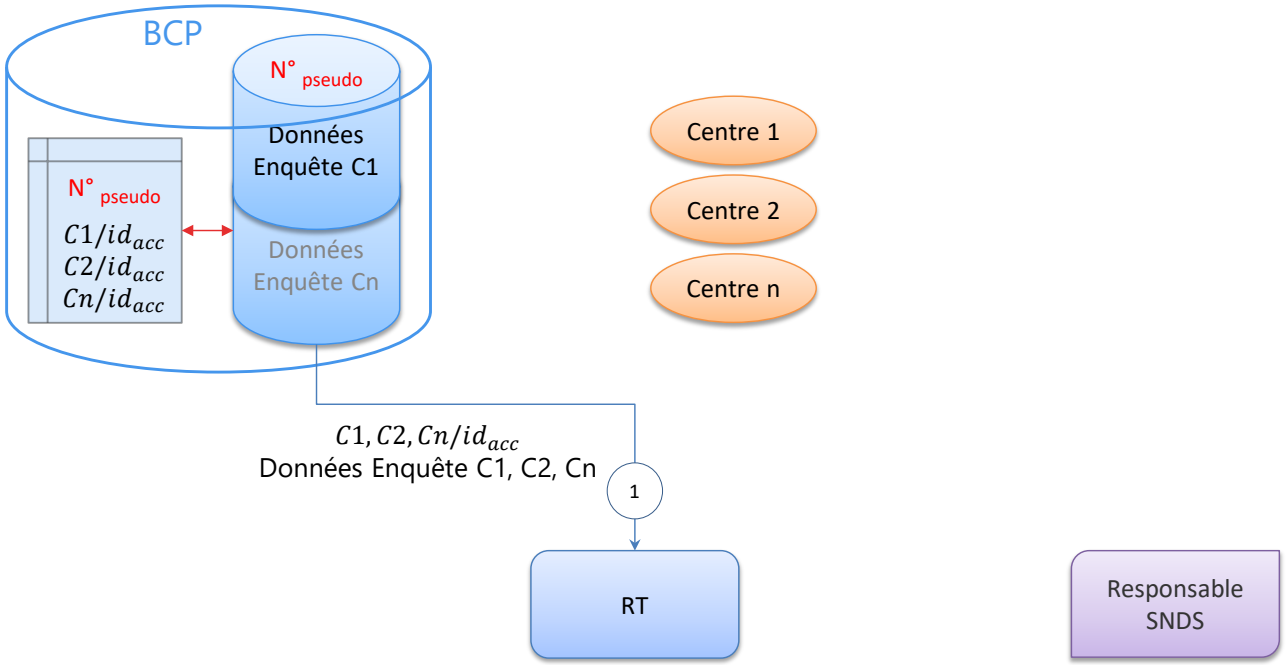
- L'utilisation d'identifiants techniques temporaires (« identifiants d'accrochage ») permet de dissocier le NIR et les données de santé lors des transferts entre acteurs.
- Par principe, **ces numéros sont non signifiants et différents à la fois du numéro d'inclusion de la personne dans l'étude et de son pseudonyme dans la base centrale**, afin de limiter les risques de réidentification croisée entre le numéro d'inclusion, le pseudonyme, le NIR et les données de santé (enquête et SNDS).
- De même, **les centres n'ont pas d'accès direct à la table de correspondance, interne à la base centrale**, qui fait le lien entre les numéros pseudonymes et les identifiants d'accrochage.
- Si le besoin est dûment justifié (par ex. transmission récurrente, audit des données, alerte des participants), la table de correspondance entre le numéro pseudonyme et l'identifiant d'accrochage peut être conservée dans la base centrale, de manière sécurisée.
- L'identifiant d'accrochage peut être généré par une fonction mathématique aléatoire, mais aussi par une fonction de hachage à clé secrète ; dans le second cas, c'est la clé secrète qui sera conservée au lieu de la table de correspondance.



Etape 1 – Extraction des données d'enquête

Les données d'enquête de l'ensemble des centres sont extraites de la base centrale, associées aux identifiants d'accrochage

- Par principe, afin de limiter les risques de réidentification portant sur les données de l'enquête et sur les données du SNDS auxquelles elles seront appariées**, le pseudonyme n'est pas extrait de la base centrale**.
- À noter : les risques de réidentification sont à considérer pour la transmission et pour le stockage des données.
- Dès lors, **pour extraire les données strictement nécessaires** (identifiants d'accrochage et données d'enquête), le RT peut demander l'intervention manuelle d'un administrateur habilité ou déclencher une fonction interne à la base.
- Dans le cas d'une fonction interne à la base, celle-ci pourra également envoyer aux centres les données qui leur sont nécessaires pour le circuit d'appariement (cf. étape 2).

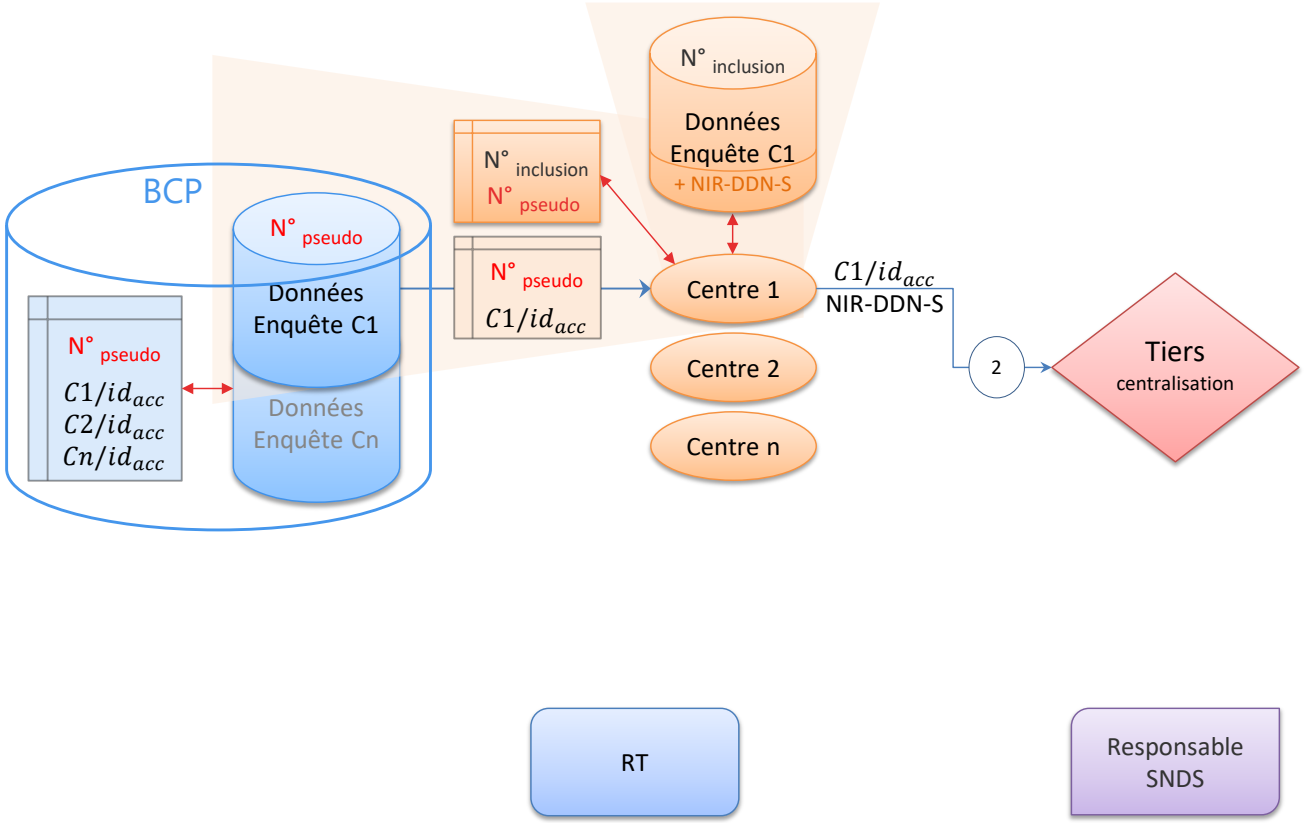


Etape 2 – Envoi des NIR au tiers centralisateur

Chaque centre reçoit la liste des pseudonymes des participants concernés, accompagnés des identifiants d'accrochage correspondants

- Par principe, **les centres n'ont pas d'accès complet à la table de correspondance de la base centrale** entre les pseudonymes et les identifiants d'accrochage. En effet, un centre ne peut être destinataire que des données qui lui sont strictement nécessaires pour le circuit d'appariement, à savoir les pseudonymes de ses seuls patients et les identifiants d'accrochage associés.

- L'extraction de ces données va donc être réalisée pour chaque centre, soit manuellement par un administrateur habilité de la base, à la demande du RT, soit par une fonction interne à la base qui sera déclenchée par le RT (cf. étape 1), soit par une fonction interne déclenchée par chaque centre dans son interface avec la base.

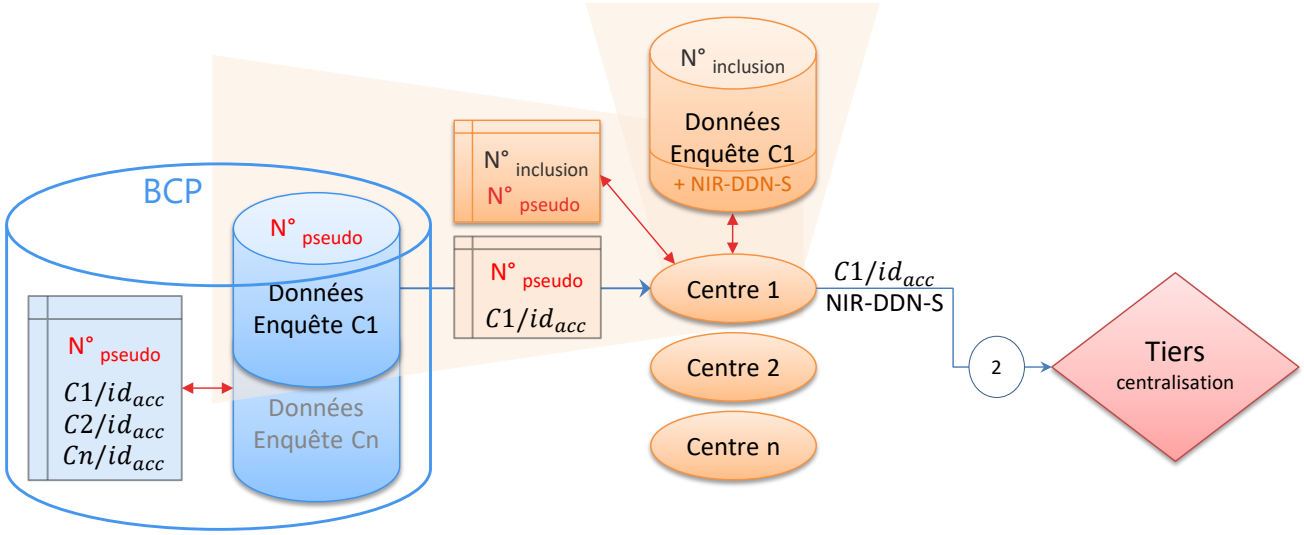


Etape 2 – Envoi des NIR au tiers centralisateur (suite)

Chaque centre transmet au tiers centralisateur les identifiants d'accrochage accompagnés des [NIR - Date de Naissance - Sexe] de ses participants

- À partir des pseudonymes reçus, à l'aide de sa table de correspondance locale avec le numéro d'inclusion et de sa base locale contenant le NIR, chaque centre va générer un fichier associant les identifiants d'accrochage reçus avec les [NIR - Date de Naissance - Sexe] correspondants. Ce fichier va être transmis au tiers centralisateur, comme seules données strictement nécessaires au circuit d'appariement.

- En effet, par principe, **le pseudonyme et le numéro d'inclusion ne sont pas transmis avec le NIR**, ces identifiants posant par nature un risque élevé de réidentification.
- De même, **le tiers n'a pas d'accès direct** aux données stockées dans les centres, et aucune autre donnée personnelle liée à l'enquête ne lui est transmise.
- Enfin, **le centre ne conserve pas le fichier généré contenant les NIR**, qui doit être détruit juste après l'envoi au tiers, **de même que celui qui contenait les pseudonymes**.



RT

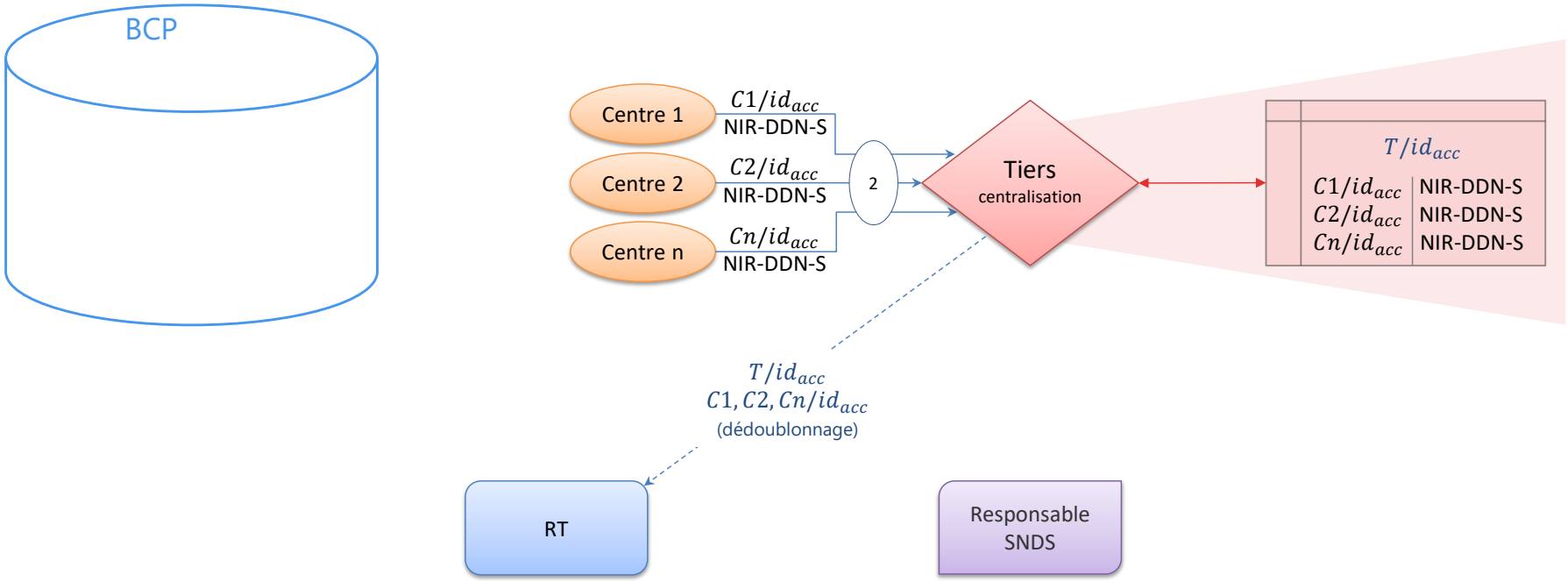
Responsable SNDS

Etape 2bis – Centralisation, dédoublonnage et envoi au RT

Le tiers centralisateur fusionne les tables reçues de chaque centre

Si nécessaire, il procède au dédoublonnage des participants ayant le même [NIR - Date de Naissance - Sexe]

- Un participant qui serait suivi par plusieurs centres aurait plusieurs numéros d'inclusion et identifiants d'accrochage associés (par ex. C1/id_{acc} et C3/id_{acc}).
- S'il est nécessaire de chaîner les données issues de plusieurs centres sur leurs participants communs, le tiers centralisateur identifie les lignes avec le même [NIR - Date de Naissance - Sexe] et leur attribue un nouvel identifiant d'accrochage unique T/id_{acc} .
- Le tiers centralisateur transmet alors au responsable de traitement la table de correspondance entre les identifiants Cn/id_{acc} et T/id_{acc} afin de permettre *in fine* d'apparier les données des centres avec celles extraites du SNDS.
- S'il n'y a pas besoin de dédoublonnage, T/id_{acc} peut reprendre les Cn/id_{acc} .

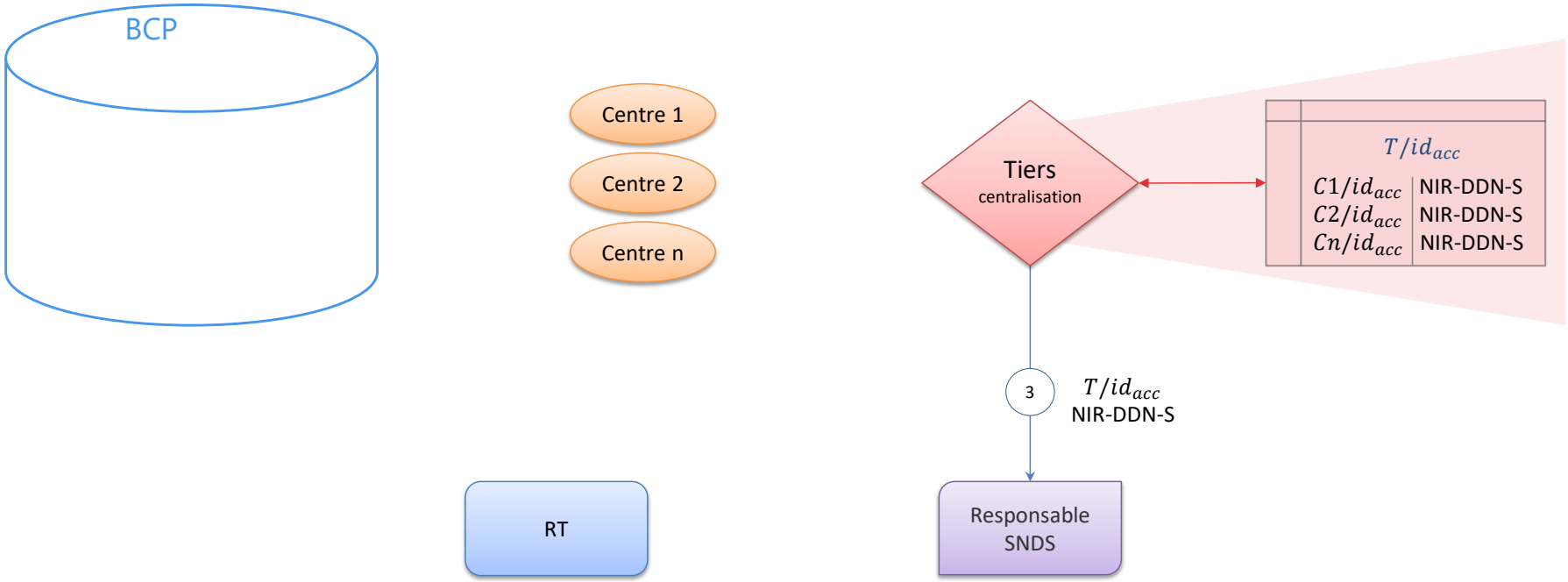


Etape 3 – Envoi des NIR au responsable SNDS

Le tiers centralisateur envoie au responsable de la base SNDS les (nouveaux) identifiants d'accrochage accompagnés des [NIR - Date de Naissance - Sexe]

- Cet envoi doit se faire à l'aide du téléservice « SAFE » de la CNAM, dans un fichier unique et au format adéquat.

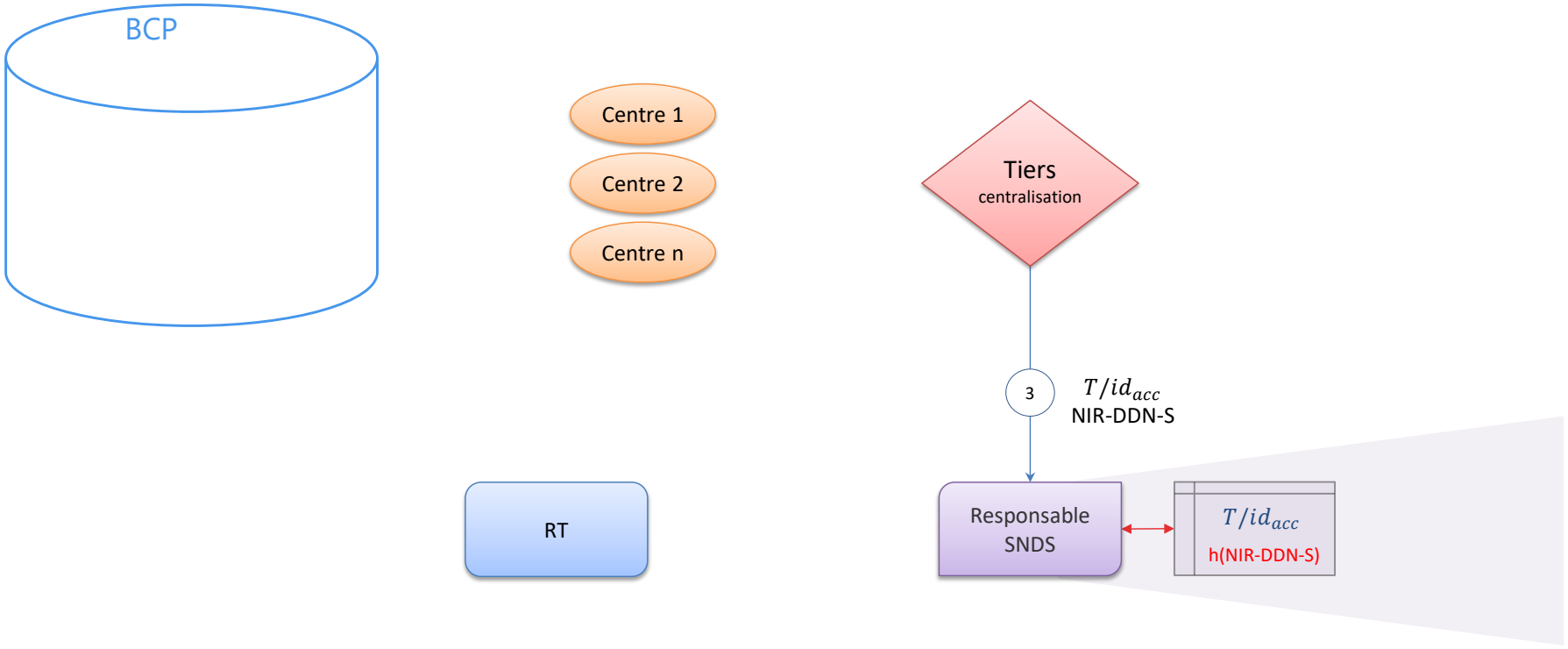
- Si le besoin est dûment justifié (par ex. transmission récurrente, audit des données, alerte des participants), la table de correspondance entre les identifiants d'accrochage Cn/id_{acc} et T/id_{acc} peut être conservée par le tiers centralisateur, de manière sécurisée.



Etape 3bis – Hachage des NIR en entrée

Dès réception, le responsable de la base SNDS procède au « hachage » du triplet [NIR + Date de Naissance + Sexe] pour générer l'identifiant interne du SNDS : h(NIR-DDN-S)

- Le hachage désigne ici un calcul cryptographique produisant une pseudonymisation irréversible.
- Dans le cas du SNDS, le NIR est pseudonymisé par plusieurs étapes de hachage successives.
- Par principe, le triplet [NIR + Date de Naissance + Sexe] est remplacé par h(NIR-DDN-S) dès réception des données et il n'est pas conservé, afin de limiter les risques de réidentification des données du SNDS.

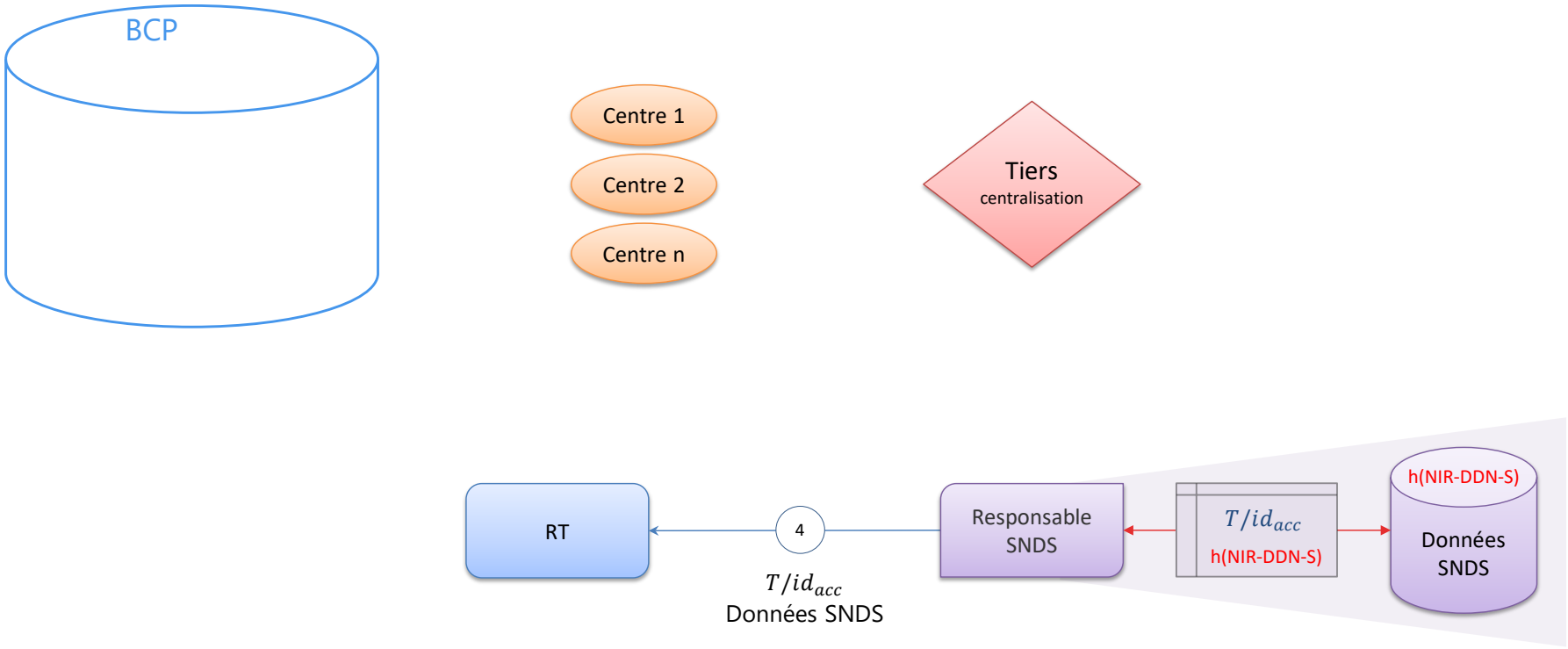


Etape 4 – Extraction et envoi des données SNDS au RT

Le responsable de la base SNDS extrait les données du SNDS correspondant aux h(NIR-DDN-S) des participants

Il transmet les données extraites au responsable de traitement, avec l'identifiant d'accrochage reçu du tiers centralisateur

- Afin de limiter les risques de réidentification des données du SNDS, son identifiant interne h(NIR-DDN-S) n'est jamais extrait : **seul l'identifiant d'accrochage est présent avec les données extraites du SND**



Etape 5 – Appariement des données d'enquête avec les données du SNDS, et génération d'un nouvel identifiant pour la base appariée

L'appariement doit être effectué sur une plateforme SNDS nationale ou dans un système (« bulle sécurisée ») conforme au référentiel de sécurité du SNDS

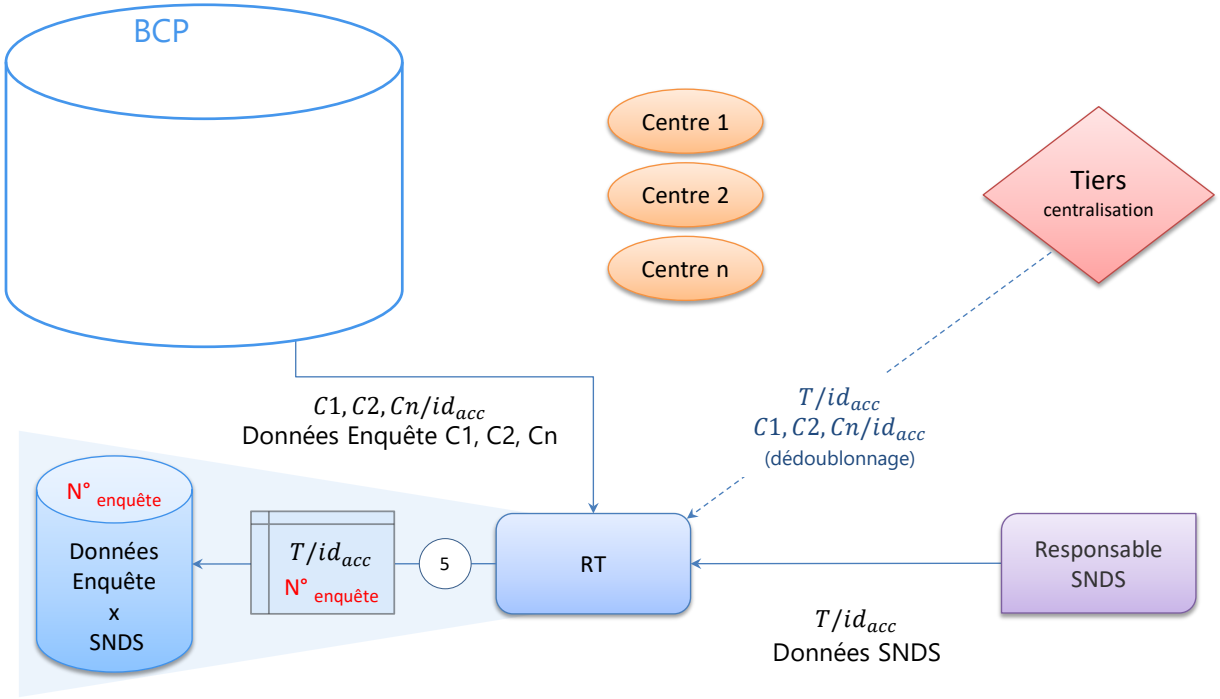
Le responsable de traitement reçoit les données du SNDS et les apparie avec les données d'enquête extraites de la base centrale, à l'aide des identifiants d'accrochage

Le cas échéant, il utilise la table de correspondance transmise par le tiers centralisateur, qui fait le lien entre les identifiants d'accrochage des centres et l'identifiant d'accrochage attribué par le tiers lors du dédoublonnage.

Après vérification de l'appariement, le responsable de traitement remplace les identifiants d'accrochage par un identifiant aléatoire propre à la base des données appariées : N° enquête

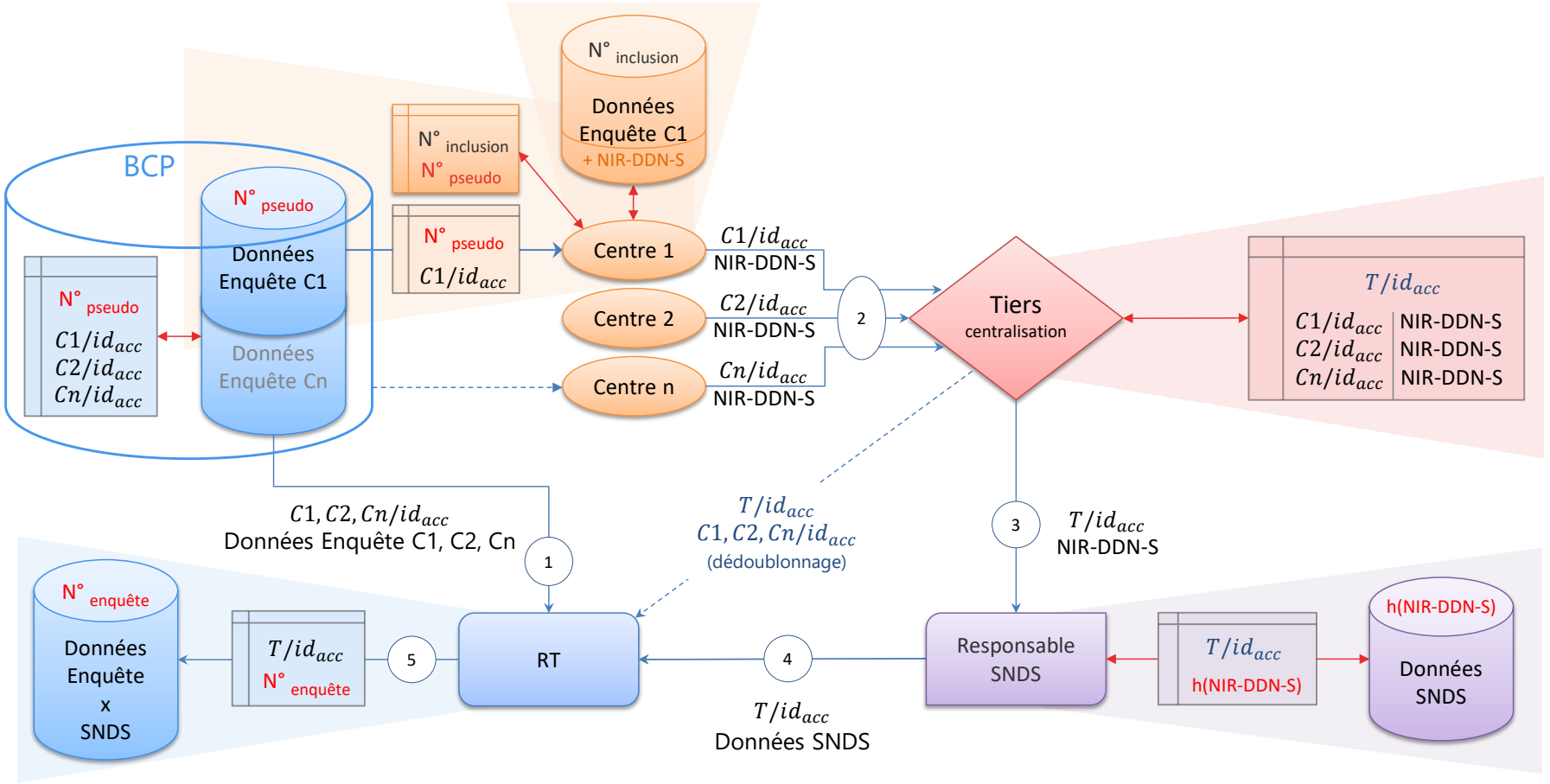
Si le besoin est dûment justifié (par ex. transmission récurrente, audit des données, alerte des participants), la table de correspondance entre les identifiants d'accrochage et l'identifiant des données appariées peut être conservée par le responsable de traitement, de manière sécurisée.

L'identifiant des données appariées peut être généré par une fonction mathématique aléatoire, mais aussi par une fonction de hachage à clé secrète ; dans le second cas, c'est la clé secrète qui sera conservée au lieu de la table de correspondance.



3. Synthèse de l'implémentation du circuit Multi-centres / Base centrale pseudonymisée

Vue fonctionnelle complète



Vue technique complète