Les Entrepôts de Données de Santé

GUIDE DE REFERENCE

BECHET Clara | Protection des données personnelles & conformité | 10/01/2024

Table des matières

Abréviations

CESREES : Comité Ethique et Scientifique pour les Recherches, les Etudes et les Evaluations dans le domaine de la Santé

CNIL : Commission Nationale de l'Informatique et des Libertés

CPP : Comité de Protection des Personnes

EDS : Entrepôt de Données de Santé (EDS)

MR : Méthodologie de référence

RGS : Référentiel Général de Sécurité

RIPH : Recherche impliquant la personne humaine

RNIPH : Recherche n'impliquant pas la personne humaine

SIH : Système d'Informations Hospitalier

SNDS : Système National des Données de Santé

Section 1 : La notion d'Entrepôt de Données de Santé (EDS)

Pour poser une définition sur la notion d'EDS, il conviendra d'étudier d'une part ses éléments de définition (§1), d'autre part ses finalités ultérieures, les projets de recherche (§2).

§1 : ELEMENTS DE DEFINITION

Un entrepôt de données se caractérise aussi bien à travers sa finalité principale qu'à travers son architecture technique, bien précise.

a) Une approche par la finalité

Pour commencer, l'entrepôt a pour objectif essentiel la collecte et la disposition de données à large échelle. Dans le domaine de la santé, cela signifie que les données primaires pourront avoir des origines variées : données issues du soin, données de recherches rétrospectives, données issues d'autre entrepôts, données issues de cohortes etc.). Généralement, l'alimentation de l'entrepôt se fait au fil de l'eau.

Ensuite, l'entrepôt a une durée de vie plus longue que les simples bases de données recherches. Plus exactement, la durée de vie d'un entrepôt est communément de vingt (20) ans maximum à compter de la collecte des données dans le cadre du soin ou de la recherche, tout dépendant de la source des données primaires et de leur cycle de vie.

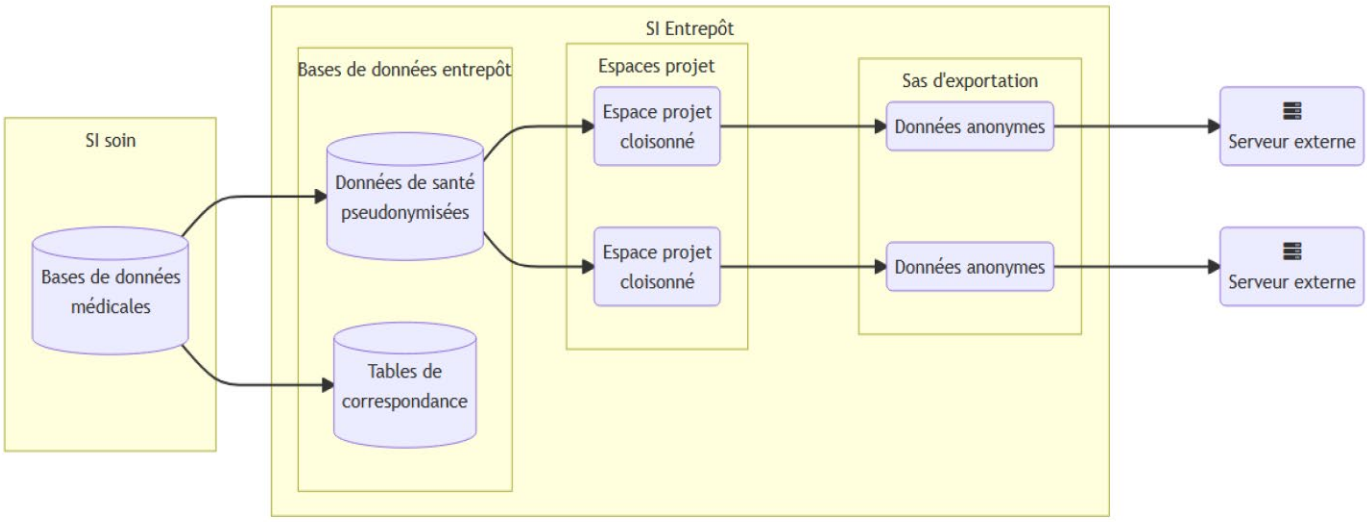
b) Une approche par l'architecture technique

Le système d'information d'un entrepôt répond à une logique et des exigences de sécurité bien précises. Pour en comprendre les raisons, il faut avoir à l'esprit qu'un entrepôt de données est un trésor intéressant pour les attaquants, non seulement parce que les reconnaissances à effectuer sont moindres mais aussi et surtout parce que les données qu'il contient sont structurées et, *de facto*, d'une grande richesse.

La création d'un entrepôt suppose nécessairement un cloisonnement réseau physique ou virtuel afin de rendre difficile l'accès au SIH à partir de l'entrepôt et réciproquement.

L'architecture du système d'information d'un entrepôt répond à un triptyque bien précis : à partir de la base de données médicales, on va pseudonymiser les données qui seront stockées séparément de la table de concordance au sein de la base « entrepôt ». A partir de cette base, des extractions seront réalisées, toujours au sein du SI entrepôt, afin de créer des espaces projets cloisonnés les uns des autres. Le cloisonnement, logique et cryptographique, suppose un chiffrement des données aussi bien au repos qu'en transit. Au repos, les exigences de l'annexe B1 du RGS doivent ainsi être respectées au niveau de la partition, ou des applicatifs, des tables de bases de données. En transit, les protocoles standard de chiffrement à jour sont à utiliser (HTTPS, TLS, SSH). Concernant le chiffrement, les données se santé, les

tables de concordances et les données génétiques le cas échéant doivent impérativement être séparées, *ie* être implémentées sur des partitions ou serveurs différents par le biais d'un chiffrement et d'un accès restreints. Enfin, la physionomie technique de l'entrepôt consiste à travailler sur des données dans des espaces projets dédiés sans qu'aucune donnée, autrement qu'anonymisée, ne puisse être exportée. Ainsi, si un projet de recherche nécessite une exportation de données pseudonymisées, une autorisation CNIL spécifique devra être obtenue pour ce projet. Les espaces de travail au sein de l'entrepôt sont cloisonnés entre eux et un nouvel identifiant par étude doit être généré afin d'éviter une fuite de l'identifiant global de l'entrepôt. Concrètement, l'accès à ces différentes bulles sécurisées passe soit par un bureau virtuel, soit par des applications mutli-*users* à condition que le cloisonnement applicatif entre les espaces demeure effectif.



Si la notion d'entrepôt peut sembler évidente, elle se distingue toutefois parfois mal des bases de données recherche qui, une fois en archivage intermédiaire, peuvent être amenées à faire l'objet de réutilisations ultérieures pour de nouveaux projets de recherche.

§2 : Distinction EDS et bases de données de recherche

Une base de données de recherche, *a contrario*, est une base qui n'est ni alimentée au fil de l'eau, ni conservée pour une longue durée.

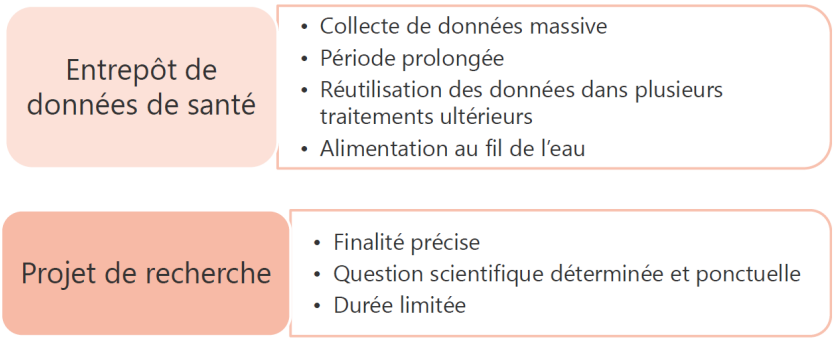
En effet, la constitution d'une base de données recherche, contrairement à l'entrepôt, a vocation à traiter les données dans le cadre d'une finalité précise, correspondant bien souvent à la question scientifique déterminée et ponctuelle posée au sein du protocole de recherche.

En France, on distingue les recherches impliquant la personne humaine des recherches n'impliquant pas la personne humaine. La première suppose une collecte des données de la recherche directement auprès des personnes concernées, la seconde suppose une collecte indirecte à partir de données déjà collectées dans le cadre du soin auprès de la personne concernée par l'étude ou dans le cadre d'une précédente recherche.

Dans ces deux types d'études, des durées de conservations limitées sont imposées en base active, *ie* pendant toute la durée du projet et la conservation des données de ces études en archivage intermédiaire[1] est elle aussi réglementée.

Si l'approche par l'architecture technique a permis de donner des éléments de définition d'un entrepôt et, *de facto*, de le distinguer d'une base de données à visée recherche. La technique de pseudonymisation utilisée permet elle aussi d'apporter des éléments de définition et de distinction supplémentaires à la caractérisation de ces deux notions. En effet, si aujourd'hui encore certaines bases de données recherche peuvent faire l'objet d'une minimisation à travers une technique de pseudonymisation « classique », comme prescrite par les méthodologies de référence recherche à savoir : utilisation des initiales (première lettre du prénom et du nom associés au mois et année de naissance), les identifiants pseudonymisés des inclusions dans l'entrepôt doivent répondre à des exigences plus poussées le rendant ainsi non devinable. Ainsi, les identifiants supposent l'utilisation d'une part de fonctions de hachage à l'état de l'art (argon2, yescrypt, scrypt, bcrypt), d'autre part d'un sel de fonctions spécialisées pour la génération de nombres aléatoires.

[1] Correspond à une séparation à la fois logique et physique de la base en archivage intermédiaire. Autrement dit, l'archivage consiste à stocker la base sur un espace de stockage différent de l'espace où se situait la base de travail avec un accès restreint. L'exemple le plus parlant est celui des contrats de travail, conservés en base active pendant toute la durée du contrat et basculés en archivage intermédiaire pendant toute la durée de la prescription en matière de contentieux prud'homal.



La distinction entre entrepôt de données et bases de recherches est importante car elle conditionne les modalités d'accès aux données et le formalisme applicable à sa constitution.

Section 2 : Le formalisme de la constitution d'un entrepôt de données de santé

La constitution d'un entrepôt de données de santé doit répondre à un formalisme bien précis (§1) tout comme tous les projets ultérieurs qui seront conduits à partir de lui (§2).

§1. LE FORMALISME LIE A L'ENTREPOT

Pour appréhender au mieux le formalisme de l'EDS il faut comprendre qu'un cadre réglementaire spécifique doit être respecté (1) ainsi que certaines exigences propres à lui (2).

a) Le cadre réglementaire

Il existe trois modalités réglementaires à respecter pour constituer un entrepôt de données de santé.

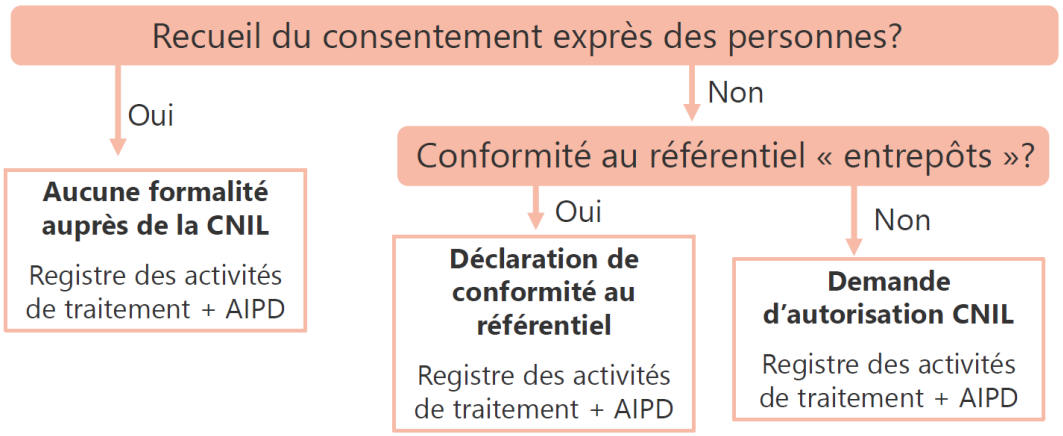
D'une part, il y a la procédure simplifiée à travers la déclaration de conformité au référentiel de la CNIL sur les entrepôts de données de santé.

D'autre part, il y a la demande d'autorisation CNIL en cas de point(s) de non-conformité au référentiel susmentionné.

La troisième modalité est le recueil du consentement exprès des personnes concernées par le pool de leurs données au sein de l'entrepôt. Pour autant, il s'agit d'une modalité réglementaire difficilement praticable sur le terrain car, bien souvent, le consentement est le levier utiliser pour faire sauter les verrous liés à certains points de non-conformité du référentiel de la CNIL. Non seulement, cette non-conformité pourrait être reprochée au responsable de traitement qui a fait passer le consentement sur le strict respect du référentiel. Surtout, la plupart des non-conformité réglementaires exposent le responsable de traitement de l'entrepôt à un risque civil et pénal vis-à-vis des personnes concernées et des fournisseurs de données.

Par conséquent, si la CNIL elle-même indique qu'aucune formalité auprès de ses instances n'est nécessaire pour formaliser la constitution d'un EDS dès lors que le consentement exprès des personnes concernées est recueilli, ce recueil ne doit pas venir supplanter pour autant le strict respect du référentiel.

Au-delà du caractère réglementaire de ce choix de conformité, la dimension politique qui s'y rattache n'en est pas moins importante et participe à la bonne image de l'institut aussi bien vis-à-vis des patients que vis-à-vis des partenaires.



b) Les principales exigences réglementaires

Base légale. Le RGPD[2] impose que chaque traitement de données à caractère personnel soit doté d'une base légale, d'un fondement juridique. La constitution d'un entrepôt de données de santé, quant à elle, est un traitement qui, conformément au référentiel EDS, doit avoir pour base légale « l'exécution d'une mission d'intérêt public ou relevant de l'exercice de l'autorité publique dont est investi le responsable de traitement. Par exemple, Gustave Roussy, en tant que Centre de Lutte Contre le Cancer et investi d'une mission publique de la recherche, peut constituer des entrepôts de données sur cette base légale.

Finalités. Le traitement des données au sein d'un entrepôt de données de santé doit nécessairement poursuivre une finalité d'intérêt public[3]. Autrement dit, chacune des finalités d'un entrepôt doit poursuivre un intérêt public et être suffisamment déterminées, explicites et légitimes[4] comme la productions d'indicateurs de pilotage, le fonctionnement d'outils d'aide au diagnostic médical, l'élaboration de projets de recherche etc. En revanche, l'entrepôt ne doit pas poursuivre les finalités interdites exposées dans le Code de la Santé Publique[5].

Données traitées. Conformément au RGPD, tout traitement de données doit respecter le principe de minimisation, *ie* ne collecter que ce qui est strictement nécessaire au regard de la finalité poursuivie par l'entrepôt. Cela signifie plus loin qu'aucune donnée ne peut être collectée au hasard et aux seuls fins d'alimenter l'entrepôt. une vigilance toute particulière doit être apportée aux données directement identifiantes, données de santé, données génétiques, le cas échéant aux données pénales, qui répondent à des exigences de sécurité précises[6]. Enfin, en cas de traitement de données du SNDS, plus exactement d'appariement aux données du SNDS, que celui-ci soit direct ou indirect, une demande d'autorisation devra être réalisée auprès de la CNIL, ce type de traitement n'entrant pas dans le champ du référentiel.

Transparence. En matière d'EDS, l'information se fait à deux niveaux : le premier à la création de l'entrepôt, le second lors de la réutilisation des données à des fins de recherches.

---

[2] Article 6-1-e) du Règlement Général sur la Protection des Données
[3] Article 66-1 de la Loi Informatique et Libertés
[4] Article 5-1-b) du Règlement Général sur la Protection des Données
[5] Article L.1461-1 V du Code de la santé publique
[6] V. Sect.1. §1 et §2

L'information, quel que soit le niveau, doit comprendre toutes les mentions obligatoires du RGPD[7] et doit être réalisée aussi bien pour les patients nouveaux qu'en cours de suivi. Une information est également obligatoire pour les professionnels de santé dont les données seront « versées » dans l'entrepôt, comme les données de connexion.

Gouvernance. Un système de gouvernance doit être mis en place afin d'avoir une parfaite maitrise de la donnée. Lors de la constitution d'un entrepôt doit être bipartite avec un comité de pilotage et un comité scientifique. Le comité de pilotage a pour objectif de déterminer les orientations stratégiques et scientifiques de l'entrepôt. Ce comité de pilotage doit également tenir une liste exhaustive des données contenues dans l'entrepôt et savoir en démontrer la nécessité. Le comité scientifique, quant à lui, est en charge de l'évaluation des demandes de réutilisation des données pour chaque projet.

§2. LE FORMALISME LIE AUX PROJETS DE RECHERCHE ULTERIEURS

On l'a compris, l'accès aux données, qu'elle qu'en soit la source : le soin, la recherche, la personne concernée, un entrepôt, est réglementé et, en fonction, les procédures d'accès ainsi que le paysage réglementaire peuvent varier.

Comme évoqué précédemment [8], il existe deux grands types de recherche, celles impliquant la personne humaine et celles n'impliquant pas la personne humaine. Classiquement, les projets de recherche conduits à partir de l'entrepôt seront des recherches n'impliquant pas la personne humaine car conduit sur des données déjà existantes, puisqu'intégrées au SI entrepôt. Pour autant, rien n'empêche la conduite d'une étude ambispective pouvant ainsi faire entrer cette dernière dans la catégorie des recherches impliquant la personne humaine. De facto, il est primordial de qualifier la recherche avant d'entreprendre les démarches adaptées au projet.

En fonction du statut juridique de l'étude, une méthodologie de référence de la CNIL devra être suivie. A défaut de conformité avec la méthodologie applicable au projet, comme par exemple, un appariement aux données du SNDS spécifique à l'étude, ou encore le recueil d'une variable non prévue par ledit référentiel, une demande d'autorisation devra être déposée auprès de la CNIL.

Ainsi, si une autorisation doit être demandée dans le cadre d'une RIPH, l'avis favorable du CPP devra être obtenue et le traitement, l'étude, autorisée par la CNIL. Si une autorisation doit être demandée dans le cadre d'une RNIPH, le dossier devra être soumis au CESREES qui transmettre à la CNIL.

[7] Articles 13 et 14 du RGPD
[8] V. Sect.1. §2

<u>Point d'attention (1)</u> : les délais d'instruction sont à prendre en compte dans la timeline du projet. Il arrive fréquemment que la CNIL proroge le délai d'instruction classique de deux mois à deux mois supplémentaires.

<u>Point d'attention (2)</u> : aucune méthodologie de référence, à ce jour, ne permet la réalisation d'une étude avec un appariement aux données du SNDS. Lorsque tel est le cas, la CNIL demande de documenter la conformité de la recherche avec la MR applicable pour la finalité hors appariement SNDS et de soumettre une demande d'autorisation de l'étude dans son ensemble, en raison de l'appariement souhaité, qu'il soit direct ou indirect.