Chapitre 3
Les données de santé

1. Introduction

Dans notre monde, qui aujourd'hui se partage entre physique et virtuel, nous assistons à la **génération d'un volume de données croissant**. Ces données, issues de sources multiples (réseaux sociaux, objets connectés, applications, photos, audio, vidéo, etc.) et aux formats variés, sont devenues essentielles au fonctionnement de notre société, voire vitales.

Il est à noter que ce contenu n'est aujourd'hui ni complètement exploité ni complètement exploitable. En fait, seule une infime partie de celui-ci est utilisée. D'après **Earthweb**, en 2022, **3 % du volume total des données produites serait à ce jour exploité**.

Les applications qui utilisent ces données sont de plus en plus nombreuses, de plus en plus efficaces et de plus en plus innovantes. Elles sont capables aujourd'hui d'utiliser et de croiser de **nombreuses sources de données** de telle sorte que nos vies deviennent de plus en plus tributaires de ces traitements et surtout **de moins en moins privées**.

Des données qui à leur origine semblaient anodines, voire anonymes, peuvent devenir après leur croisement avec d'autres sources, particulièrement sensibles. C'est notamment le cas pour les données de santé.

Les données nous permettent de choisir de prendre des décisions, elles nous facilitent la vie mais pas uniquement, elles nous sont aussi vitales. Que deviendrions-nous sans elles ? Imaginez en vrac : les comptes bancaires, la fraude, la sécurisation de sites sensibles et des prisons, les réseaux sociaux, les documents dématérialisés, la gestion de l'électricité, les moyens de transport, les moyens de navigation dont le GPS, les voyages, les réservations de toutes sortes, la gestion des hôpitaux et des dossiers patients, la prise de rendez-vous, les prothèses intelligentes, etc.

Dans la suite de ce chapitre, après avoir détaillé **les concepts et problématiques généraux liés aux données**, je vais mettre **l'accent sur la gestion, la conservation, l'utilisation et la sécurisation des données** qui sont issues et utilisées par le domaine de la santé.

Ce chapitre s'adresse avant tout à la génération et à l'utilisation des données de santé en France, je ferai néanmoins référence à d'autres pays lorsque cela sera nécessaire.

2. Le type de données

Les données sont scindées en **trois types distincts** :

- **Les données structurées** : elles résident dans des formats et des modèles prédéfinis.
- **Les données non structurées** : elles sont stockées dans leur format naturel jusqu'à ce qu'elles soient extraites pour analyse.
- **Les données semi-structurées** : elles sont essentiellement un mélange de données structurées et non structurées.

Un quatrième type de données est aujourd'hui à considérer : **les données synthétiques**. Néanmoins, comme nous le verrons dans la suite de ce document, les données synthétiques peuvent être vues comme un sous-ensemble des types précédents.

2.1 Les données structurées

Pour la plus grande part, **les données structurées sont des informations qui sont organisées dans des fichiers texte, des feuilles de calcul et des bases de données relationnelles**. On y trouve : les numéros de téléphone, de sécurité sociale, les divers codes, les champs textuels fixes ou variables dans les enregistrements des bases de données, des ensembles de nombres, etc.

Ces données peuvent être **générées par un humain ou par une machine**, mais la condition première est qu'elles puissent être facilement **classées**, **retrouvées** et **extraites**.

Les entreprises et organismes sont familiarisés avec ce type de données, que l'on considère comme des données ou des sources de données « **traditionnelles** ».

Ces données sont généralement stockées dans des systèmes de bases de données, le plus souvent relationnelles, et associées au système « *Back-end* » de l'entreprise qui comprend les solutions de *CRM : Customer Relationship Management, ERP : Enterprise Resource Planning, SCM : Supply Chain Management, BI : Business Intelligence and Data Mining*, et aux applications spécifiques au secteur ou à l'entreprise.

La particularité des données structurées est que leurs volume, durée de vie et croissance sont connus et peuvent être anticipés.

2.2 Les données non structurées

Les données non structurées ont une structure interne, mais celle-ci ne suit pas les modèles de données ou schémas prédéfinis. Ces données sont très variées, elles peuvent être textuelles ou non textuelles, générées par un humain, une machine ou un objet et peuvent être stockées dans des bases de données spécifiques aux données non structurées.

2.2.1 Exemples de données non structurées générées par des humains

- **Les fichiers texte** : documents créés à l'aide d'un traitement de texte, feuilles de calcul, présentations, logs (fichiers de trace).
- **Les courriers électroniques** : on les considère parfois comme des données semi-structurées, car on leur associe des métadonnées. Mais, dans tous les cas, le contenu d'un courrier électronique est non structuré puisqu'il ne peut pas être traité par des outils d'analyse classiques.
- **Les réseaux sociaux** : ce sont les données en provenance des réseaux privés ou des réseaux professionnels, tels que Facebook, X, LinkedIn, TikTok, etc.
- **Les sites web spécialisés** : YouTube, Instagram, Flickr et les sites de partages de contenus (photos, vidéos, jeux, sons, musiques, etc.).
- **Les données en provenance des téléphones portables et tablette** : messages, positionnements, comportements, habitudes d'achat, sites visités, contacts, etc.
- **Les communications instantanées ou pas** : chats, IM, enregistrements téléphoniques, outils de collaboration.
- **Les divers médias** : MP3, photos digitales, fichiers audio et vidéo.
- **Les applications métier** : documents, notes, etc.

2.2.2 Exemples de données non structurées générées par des machines ou des objets

- **Les images satellites** : données météo, paysages, mouvements militaires ou autres.
- **Les données scientifiques** : exploration pétrolière, exploration spatiale, imagerie sismique, données atmosphériques.
- **La surveillance digitale** : photos ou vidéos de surveillance.
- Les données en provenance des **capteurs et des sondes** : trafic, climat, géographie, centre de données, véhicules, etc.

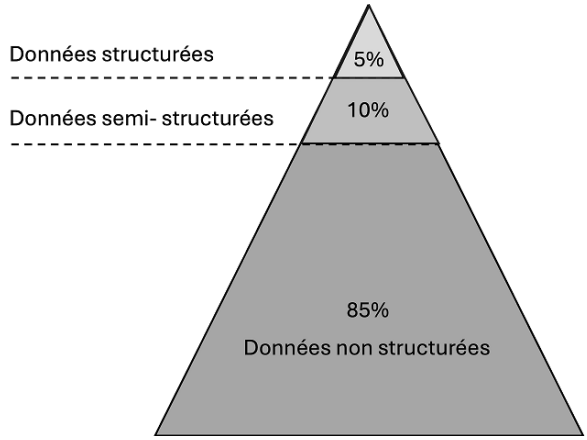
Un nombre important de données non structurées est produit par des machines ou des objets. L'humain n'ayant pas obligatoirement connaissance de celles-ci, on parlera dans ce cas de données cachées (Hidden Data). Les données cachées représentent **60 % des données non structurées**.

2.3 Les données semi-structurées

Les données semi-structurées sont une forme de données structurées qui n'obéissent pas à la structure des modèles de données associés aux bases de données relationnelles ou à d'autres formes de tables de données. Néanmoins, elles contiennent une certaine structure, notamment grâce aux **balises** ou autres **marqueurs** qui permettent de séparer les différents éléments sémantiques. On parle également de **structure auto-descriptive**.

Même si les données ne sont pas totalement structurées, **la structure auto-descriptive permet d'analyser plus facilement les données**. Elle peut aussi, au travers de processus et développements assez simples, permettre le stockage de ces données dans des bases de données relationnelles.

Le stockage et l'utilisation des données semi-structurées est moins facile qu'avec des données structurées, mais plus simple qu'avec des données non structurées. Il faut noter que l'interprétation de relations entre les données semi-structurées est rendu difficile par leur structure irrégulière et partielle et par le fait que certaines sources ont une structure de données implicite.



Données structurées, semi-structurées et non structurées

L'accroissement exponentiel du volume de données produit, permet d'avancer le nombre de **2142 Zo (zettabytes, soit 2142 suivi de 21 zéros) comme étant le volume global** des données qui sera stocké un peu partout dans le monde et sur tout type d'objet en 2035.

Il sera impossible dès lors de faire transiter ces données sur les réseaux existants, quelles que soient les technologies employées. Cela signifie que ces données devront être stockées différemment et, pour la majorité d'entre elles, ne pourront pas bouger. Elles devront rester et être traitées là où elles sont produites, aussi la sécurité de ces données sera un enjeu essentiel.

2.4 Les données synthétiques

Les données synthétiques (« synthetic data ») sont des données générées artificiellement via des algorithmes informatiques.

Les jeux de données synthétiques peuvent être de **deux formes** soit **totalement synthétiques**, **soit partiellement synthétiques**

créées en complément d'un jeu de données réelles (« *Data Augmentation* ») ou en remplacement de données réelles sensibles, par exemple **l'anonymisation** (« *Data Alteration* »).

Les données synthétiques dans le domaine de la santé offrent plusieurs avantages, notamment en matière de préservation de la vie privée en générant des données qui imitent les informations réelles des individus mais ne font référence à aucun humain.

3. Les données de santé

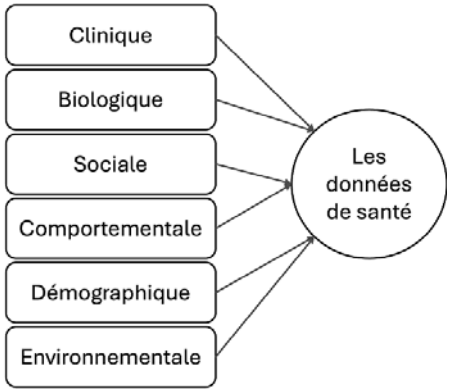
3.1 Définition

Les données de santé sont des données à caractère personnel et considérées comme sensibles. Elles font à ce titre, l'objet d'une protection particulière par les textes (règlement européen sur la protection des données personnelles, loi Informatique et Libertés, code de la santé publique, etc.). **Le règlement européen** sur la protection des données personnelles (**RGPD**) qui est entré en application le 25 mai 2018, **définit les données de santé comme** :

- **Les informations relatives à une personne physique**, collectées lors de son inscription en vue de bénéficier de services de soins de santé ou lors de la prestation de ces services.
- **Les informations obtenues lors du test ou de l'examen** d'une partie du corps ou d'une substance corporelle, y compris à partir des données génétiques et échantillons biologiques.
- **Les informations concernant** une **maladie**, un **handicap**, un **risque de maladie**, les **antécédents médicaux**, un **traitement clinique** ou **l'état physiologique** ou **biomédical**, de la personne concernée indépendamment de sa source.
- **Toutes données de mesure** à partir desquelles il est possible de déduire une information sur l'état de santé de la personne.

Il faut être conscient que le croisement des données de santé avec d'autres données peut permettre de tirer une conclusion sur l'état de santé ou le risque pour la santé d'une personne et peut ainsi générer des données de santé sensibles.

Il est à noter que la **loi ne s'applique pas** aux traitements qui comporteraient des données de santé à l'usage exclusif de la personne. À titre d'exemple, la loi ne s'applique pas aux applications mobiles en santé qui proposent dans leurs fonctionnalités, la collecte, l'enregistrement ou la conservation de données, à condition que **ces opérations s'effectuent localement** sur un ordinateur, un smartphone ou une tablette, sans connexion extérieure et à des fins exclusivement personnelles.



Les sources des données de santé

Une fois **la qualification de données de santé retenue**, un régime juridique particulier justifié par la sensibilité des données va s'appliquer. La liste suivante n'est pas exhaustive, mais permet de se rendre compte des **législations susceptibles de s'appliquer** :

- La loi Informatique et Libertés (art. 8 et chapitre IX (articles 53 à 61)).
- Les dispositions sur le secret (art. L. 1110-4 du CSP).
- Les dispositions relatives aux référentiels de sécurité et d'interopérabilité des données de santé (art. L. 1110-4-1 du CSP).
- Les dispositions sur l'hébergement des données de santé (art. L. 1111-8 et R. 1111-8-8 et s. du CSP).
- Les dispositions sur la mise à disposition des données de santé (art. L. 1460-1 et s. du CSP).
- L'interdiction de procéder à une cession ou à une exploitation commerciale des données de santé (art. L. 1111-8 du CSP, art. L 4113-7 du CSP).

3.2 Stockage et sécurisation des données de santé

Les professionnels de la santé tels que les médecins, les hôpitaux, les organismes sociaux, la médecine du travail et autres, ont besoin de stocker les données récoltées dans de bonnes conditions et de manière sécurisée.

Nous l'avons vu et vécu partout dans le monde ces dernières années, **les données de santé sont particulièrement exposées aux risques de piratages informatiques**, avec des conséquences souvent importantes et médiatisées. Dans ce contexte, la gestion des données de santé depuis leur collecte, traitement, archivage jusqu'à une éventuelle diffusion demande un besoin accru de confidentialité et de protection.

En France, depuis 2018, la **certification** « **Hébergement de données de santé (HDS)** » existe pour cela. Cette certification est obligatoire pour tous les acteurs qui recueillent des données de santé à caractère personnel, quelles que soient leurs origines. L'objectif principal de cette certification est de garantir la confiance dans les tiers auxquels des structures et des professionnels des secteurs de la santé confient les données de santé qu'ils produisent ou recueillent sur supports numériques.

La certification HDS