# De la collecte des données de santé

Création **les Sismo**

## L'usage primaire de la donnée

**Compte-rendus d'examen, résultats biologiques, prescriptions de médicaments, etc. Derrière chacun de ces documents se cachent des étalines de données. Ces dernières sont essentielles à la prise en charge du patient.**

**Elles soutiennent le personnel hospitalier dans le diagnostic, la prescription, et permettent un meilleur suivi thérapeutique.**

Chaque jour,
**22 000**
patients sont accueillis au sein des 38 hôpitaux de l'AP-HP

eds Usbome le potentiel de nos données de santé
AP-HP

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS
GREATER PARIS UNIVERSITY HOSPITALS

# 1 Sélection et extraction de la donnée brute

Dans le respect du RGPD, les données d'intérêts sont ciblées, copiées et documentées dans une première base de stockage propre à l'AP-HP (appelée base d'intégration) sans transformation à ce stade. Cela permet de regrouper dans un même espace de traitement ces données initialement éparpillées.

**Une bonne extraction prépare la suite :** si l'extraction est mal définie, tout le reste (transformation, sécurité, analyse) devient plus long et plus risqué.

### Une documentation rigoureuse des données extraites

On doit savoir d'où vient la donnée, qui l'a saisie, et quand. Les données extraites font donc l'objet d'une documentation afin d'indiquer le contexte de leur création, alimentant le travail d'interprétation à venir.

### Création de la donnée

Les données sont renseignées dans les logiciels dédiés.

Il existe plus de **350** logiciels recensés au sein de l'AP-HP

ORBIS
GLIMS
Imagerie
ODOC

### L'interopérabilité des données face à la diversité des systèmes d'information

La multiplicité des systèmes informatiques limite la transversalité des données. Assurer l'interopérabilité est une condition *sine qua non* pour permettre la coopération de systèmes hétérogènes, garantissant que la donnée partagée reste intelligible, fiable et immédiatement opérationnelle pour l'utilisateur final.

### Une extraction parcimonieuse et réfléchie

On ne peut pas tout extraire, tout le temps, de la même manière, car cela a un coût financier et énergétique très conséquent.
Des choix d'intégration de données à l'Entrepôt sont faits : périmètre, fréquence, formats, priorités.

### Explorer les données

La data discovery est le processus d'exploration, de cartographie et de compréhension globale des données d'une organisation. Son but est simple : savoir quelles données existent, où elles dorment, ce qu'elles signifient et comment les exploiter. En bref, c'est passer d'une masse de données inconnues à un patrimoine cartographié et valorisé.

Cochin - diabétologie · 16.03.2026 · Service · Tom Ardisi · masculin · antipyrétique · Identité · Sexe · Symptôme · 42 kg · 160 cm · 200/90 mmHg · hypercalcémie · Poids · Taille · Tension · Diagnostic · metformine · 2000 mg · Traitement · Image

### Des données multiples et hétérogènes

Les données de santé sont **produites dans des contextes cliniques, organisationnels et techniques très variés.**
Cette diversité conduit à des représentations multiples d'une même réalité médicale.

1. **Les données sont enregistrées dans des systèmes hétérogènes reposant sur des modèles propres à chaque éditeur ou domaine métier :** texte libre, données structurées, terminologies, documents, signaux, images, etc.
2. **Leur granularité et leur niveau de structuration varient selon le contexte de production :** urgence, suivi longitudinal, spécialité médicale, recherche, pilotage, facturation, etc.
3. **Leur interprétation dépend fortement de leur contexte de production, de leur temporalité, de leur provenance et du niveau de validation clinique associé.**

# 2 Mise en qualité de la donnée

La mise en qualité de la donnée constitue un levier essentiel pour garantir la valeur, la fiabilité et la réutilisabilité des informations produites au sein du système d'information. Dans un environnement marqué par la diversité des applications, des formats et des usages, la mission consiste à structurer, standardiser et contrôler les données afin de favoriser leur compréhension partagée, leur circulation et leur exploitation sécurisée.

Stockage de la donnée extraite

Données brutes

### Enrichir

**Les datascientists** conçoivent, développent et maintiennent les modèles d'apprentissage et bibliothèques de données afin :

- d'extraire les informations clefs à partir de données non structurées (ex : des textes dans les documents médicaux)
- de faciliter l'exploitation des données une fois enrichies.

Conception de modèles d'apprentissage capables de traiter des lots de données

Contact des médecins pour comprendre le contexte de production de la donnée

### Comprendre et documenter

**Les data-analysten** étudient les données afin de bien les comprendre et les documenter dans le catalogue de données.

- Ils contrôlent la qualité des données et documentent leur fiabilité: données manquantes, valeurs aberrantes, fréquence de mise à jour.
- Ils relient les concepts médicaux aux tables et colonnes des bases de données hospitalières, par exemple "tension artérielle systolique" ou "date d'entrée en réanimation".
- Ils cartographient la gouvernance en identifiant les responsables médicaux et techniques, en lien avec le Délégué à la Protection des Données associé.

Ils retracent et cartographient le parcours des données, de leur source médicale jusqu'à leurs usages.

Ils documentent et explicitent les concepts cliniques et administratifs au sein du Glossaire pour s'assurer que soignants, chercheurs et gestionnaires partagent les mêmes définitions (ex : ce qu'on entend par "séjour" ou "comorbidité")

Ils identifient les données sensibles et leur niveau de protection, dans le respect du RGPD du secret médical. Ils précisent les règles d'accès, les usages autorisés et les modalités de consentement des patients.

### Suivre la préparation

Toutes les étapes de préparation de la donnée dans la Base Centrale sont coordonnées par les **data product owners** qui assurent le suivi et la qualité de la donnée pour que celle-ci puisse servir les besoins initiaux de recherche et de pilotage identifiés.

Données exploitables

# 3 Constitution d'une base de données prêtes à l'emploi

Après toutes ces transformations, on obtient un ensemble de jeux de données cohérents et juridiquement conformes (selon la loi Informatique et Liberté de la CNIL et le Code de la Santé Publique), prêts à être mis à disposition des projets de recherche et de pilotage.

### Mettre à disposition

Toutes les données de santé disponibles dans la base centrale sont classées par type et exposées dans un catalogue, prêtes à être mises à disposition.

### Standardiser

**Les data-ingénieurs** appliquent le modèle de structuration de la donnée en versant les données extraites dans les modèles définis :

- pour standardiser les formats d'informations selon les règles prédéfinies et les nomenclatures internationales (FHIR) ;
- pour supprimer les doublons d'informations ;
- pour chainer entre elles les données reliées (même parcours patient, même service, ...).

Liste des patients sous anti-infectieux

Niveau de détail : UF

### Sécurité et gouvernance de l'offre de services

La mise à disposition des données dans une base centrale ne consiste pas en un simple stockage. Elle repose sur un environnement sécurisé et sur une gouvernance définie, qui encadre les conditions d'accès, d'utilisation et de suivi des données.

Autrement dit, la base centrale constitue un cadre organisé, fondé sur des règles d'accès, des mécanismes de traçabilité et une documentation permettant de comprendre et de réutiliser les données dans de bonnes conditions.

## Base centrale

**+ de 43 millions** d'examens d'imagerie
**+ de 214 millions** de compte-rendus médicaux
**+ de 1,44 milliards** de résultats de biologie

Dossiers de spécialités Orbis®
Imagerie médicale
Documents médicaux
Biologie
Immunologie
Transcriptomique
Niveau de détail : UF
Mesures cliniques
Médicaments
Médico-administratif
Socio-démographie des patients

### Une mise à disposition encadrée et sécurisée

La sensibilité des données de santé impose des mesures de protection spécifiques. Des dispositifs techniques et organisationnels sont mis en place pour prévenir l'identification directe des personnes, limiter l'accès aux seules données nécessaires et encadrer strictement leurs usages.
Ces dispositifs participent ainsi à sécuriser les traitements menés à des fins de recherche.

La base centrale contient les données de plus de
**20 millions** de patients

### Pilotage de l'hôpital

Les données de santé peuvent servir à suivre et analyser l'activité hospitalière. Elles permettent de produire des indicateurs pour mieux comprendre les pratiques et guider les décisions d'organisation. Par exemple, dans les maternités elles peuvent servir à suivre l'activité, faire le diagnostic des pratiques et mettre en œuvre des actions correctives.

### Soutien de la recherche

Plus de 350 projets de recherches multicentriques sont menés à l'appui de l'EDS et montre comment l'exploitation des données de santé permet de faire avancer la recherche.

Par exemple, dans le projet ODDIA, l'exploitation des données de 1.3 million de patients diabétiques a permis aux chercheurs de mieux comprendre les risques et confirmer des résultats à grande échelle. Ces données de "vie réelle" offrent une vision plus complète des patients et soutiennent l'amélioration des connaissances médicales.

### Développement des outils d'IA

Le travail sur la donnée est essentiel pour développer des outils d'intelligence artificielle comme Chat DPI, capables de retrouver rapidement l'information pertinente dans les dossiers patients, de la synthétiser et d'assister les soignants dans leurs décisions.

### Accompagnement des patients

Les données de santé permettent aussi de mesurer et d'améliorer la qualité des soins. Par exemple, à partir des lettres de sortie de patients, il est possible de produire des indicateurs automatiques pour suivre la complétude des documents, les délais de validation ou encore comparer les pratiques entre services.

Ces analyses aident à objectiver les écarts et à mettre en place des actions d'amélioration ciblées au bénéfice des patients.

## L'usage secondaire de la donnée

**Les données peuvent être réutilisées au-delà de la prise en charge et du suivi du patient. Après transformation, elles permettent de produire des connaissances utiles au collectif, qu'il s'agisse de faire avancer la recherche, d'améliorer les soins ou d'appuyer le pilotage de l'activité hospitalière.**

# ... à leur exploitation

Pour d'infos sur l'Entrepôt des Données de Santé