Une image contenant Graphique, logo, graphisme, Police

Description générée automatiquementUne image contenant texte, clipart, Police, logo

Description générée automatiquementUniversité Abdelmalek Essaadi

Faculté des Sciences et Techniques de Tanger

Département Génie Informatique

**Projet de Fin de Module - Data space, integration & semantic operability**

**European Health Data Space (EHDS / HDS)**

**Une image contenant symbole, Police, Graphique, capture d’écran

Description générée automatiquement**

Réalisateurs :

Imad El Maftouhi, Diae Khayati, Hamza Ait Youssef,

Amina Louazir, Salma Maaquili

Supervisée par : Pr. Baïda Ouafae

MST Intelligence Artificielle et Science des Données

M233 – Data space, integration & semantic operability

**2025/2026**

# Abstract

L’Espace européen des données de santé (European Health Data Space – EHDS) constitue un pilier fondamental de la stratégie européenne des données, visant à permettre le partage sécurisé, interopérable et souverain des données de santé à l’échelle de l’Union européenne. Malgré un cadre réglementaire ambitieux, la mise en œuvre concrète de l’EHDS soulève des défis techniques majeurs, notamment liés à l’intégration de sources hétérogènes, à l’interopérabilité sémantique, à la qualité des données et au respect strict des exigences de protection des données personnelles.  
  
Ce projet propose une étude scientifique et expérimentale du EHDS à travers la conception et l’implémentation d’un prototype reproductible illustrant les mécanismes d’intégration des données et d’interopérabilité sémantique. Trois familles de sources de données de santé sont intégrées : des dossiers médicaux électroniques synthétiques, des données cliniques destinées à l’usage secondaire, et des données de laboratoire hétérogènes. Un pipeline ETL implémenté en Python assure l’extraction, la transformation, la pseudonymisation et l’harmonisation des données, produisant une base intégrée exploitable.  
  
Pour répondre aux problématiques sémantiques, un modèle basé sur les standards du Web sémantique (RDF, SKOS et OWL) est conçu. Les données intégrées sont représentées sous forme de graphe de connaissances, permettant l’exécution de requêtes SPARQL avancées et la démonstration de raisonnements inter-sources, notamment pour la détection de contre-indications médicales. L’ensemble du travail est conçu pour être auto-suffisant, reproductible et aligné sur les exigences du règlement EHDS.

# PARTIE I – Étude approfondie du Data Space

Cette section vise à fournir une analyse complète et structurée de l'Espace européen des données de santé (EHDS). Elle établit les fondements conceptuels, organisationnels et techniques du projet, en mettant l'accent sur les défis liés à l'intégration et au partage des données de santé entre des systèmes de soins hétérogènes, tout en garantissant l'interopérabilité, la sécurité et la conformité réglementaire au niveau européen.

## Présentation générale

Cette section présente l’Espace européen des données de santé (EEDS) et le situe au sein du secteur des soins de santé et de la santé publique. Elle décrit le contexte réglementaire et technologique dans lequel l’EEDS opère, notamment la nécessité d’un accès transfrontalier aux données de santé au sein de l’Union européenne. Les objectifs de l’EEDS sont présentés, couvrant à la fois l’utilisation primaire des données pour les soins aux patients et leur utilisation secondaire pour la recherche, les politiques de santé publique et l’innovation. Une attention particulière est portée aux défis commerciaux et techniques tels que la fragmentation des systèmes d’information de santé, l’hétérogénéité des normes de données et les exigences strictes en matière de confidentialité, de sécurité et de protection des données.

### Nom du Data Space

**Espace européen des données de santé (EHDS)**  
**(Règlement (UE) 2025/327)** : Premier espace de données sectoriel de la stratégie européenne des données, créant un écosystème unifié pour la gestion des données de santé électroniques au sein de l'UE.

### Secteur d’activité

* **Secteur principal** : Santé (soins de santé, santé publique)
* **Secteurs connexes** :
  + Recherche médicale et scientifique
  + Industrie pharmaceutique
  + Dispositifs médicaux et intelligence artificielle en santé
  + Innovation numérique en santé
  + Élaboration des politiques de santé publique
* **Intégration** : Union européenne de la santé (résilience face aux pandémies, vieillissement, pénuries de main-d'œuvre)
* **Portée géographique** : 27 États membres de l'UE, avec possibilité d'extension à des pays tiers

### Contexte et objectifs

L’Espace européen des données de santé (EHDS) est une initiative centrale de la stratégie européenne pour les données de 2020. Il répond à la fragmentation historique des systèmes de santé en Europe, qui entrave les soins transfrontaliers et limite la recherche. La pandémie de COVID-19 a accentué la nécessité d’un partage de données rapide et sécurisé.

Le règlement, proposé par la Commission le 3 mars 2022 après de larges consultations, a été adopté le 11 février 2025 et est entré en vigueur le 26 mars 2025. Il s’intègre aux cadres existants (RGPD, Règlement sur l’IA, etc.) pour garantir sécurité et éthique.

L’EHDS poursuit un double objectif. D’une part, l’**utilisation primaire** donne aux citoyens le contrôle et la portabilité de leurs données de santé électroniques, facilitant les soins à travers l’Europe (ex. : ordonnances électroniques valables à l’étranger). D’autre part, l’**utilisation secondaire** permet la réutilisation sécurisée et strictement encadrée de données pseudonymisées pour la recherche, l’innovation et les politiques de santé publique, via l’infrastructure **HealthData@EU**.

L’initiative vise aussi à créer un marché unique des solutions de santé numériques par la certification des systèmes, ce qui devrait générer d’importantes économies et stimuler la croissance du secteur. Sa mise en œuvre est progressive, avec des échéances étalées jusqu’en 2031, et s’appuie dès à présent sur un vaste portefeuille de projets préparatoires financés par l’UE.

### Enjeux métiers et techniques

L’EHDS repose sur une architecture technique fondée sur l’**interopérabilité** au moyen de normes européennes (EEHRxF, FHIR, SNOMED CT, ICD-11). Cette approche structure le partage autour de silos nationaux fédérés via **HealthData@EU**, avec un traitement sécurisé dans des environnements contrôlés et un accès via des catalogues et API.

La **gouvernance** associe des organismes nationaux d’accès (HDAB), un Conseil EHDS de coordination et des règles strictes de protection (RGPD), incluant pseudonymisation et anonymisation. Les principaux défis concernent la **confidentialité** (opt-out des patients, prévention des utilisations abusives), les **écarts de maturité numérique** entre États membres, la **qualité des données** et les délais de transposition des actes d’exécution.

Des projets en cours (plateforme HealthData@EU 4.0, SHAIPED) et des approches comme l’**apprentissage fédéré** visent à répondre à ces enjeux. La réussite de l’EHDS dépendra également de la prise en compte des dimensions **éthiques et sociales** (équité, non-discrimination, transparence), nécessitant une collaboration soutenue entre toutes les parties prenantes pour consolider la position de l’UE comme leader en santé numérique.

## Acteurs et gouvernance



Figure 1 : coalition des parties prenantes de l'EHDS

Cette section présente les principaux acteurs de l’écosystème EHDS : les **fournisseurs de données** (hôpitaux, autorités sanitaires), les **utilisateurs** (soignants, chercheurs) et les **autorités de confiance** (points de contact nationaux, HDAB). Elle décrit également le modèle de gouvernance qui encadre l'accès, le consentement et le partage des données selon le RGPD et le règlement EHDS.

### Fournisseurs de données

Les fournisseurs (data holders) sont les entités qui produisent ou détiennent des données de santé électroniques. Il s'agit notamment des hôpitaux, cliniques, laboratoires, pharmacies et autorités sanitaires, ainsi que des fabricants de dispositifs médicaux et des éditeurs de logiciels. Dans le cadre de l'EHDS, ils ont l'obligation de mettre à disposition les données pour l'usage primaire et, sous conditions, pour l'usage secondaire. Des exceptions existent pour les micro-entreprises ou la protection des secrets commerciaux. Ils peuvent facturer des frais raisonnables pour la préparation des données et signaler les risques de réutilisation abusive.

### Consommateurs de données

Les utilisateurs (data users) diffèrent selon la finalité. Pour l'usage primaire, ce sont principalement les **patients**, qui peuvent contrôler et partager leurs données via MyHealth@EU, et les **professionnels de santé** pour assurer la continuité des soins transfrontaliers. Pour l'usage secondaire, il s'agit des **chercheurs**, des **industries pharmaceutiques**, des **autorités de régulation**, des **innovateurs** (IA) et des **décideurs politiques**. L'accès aux données secondaires est strictement conditionné par un permis délivré par les HDAB et s'effectue uniquement dans des environnements de traitement sécurisés. Les organisations de patients jouent un rôle clé dans la légitimation éthique de ces usages.

### Intermédiaires / autorités de confiance

Les principaux intermédiaires sont les **Health Data Access Bodies (HDAB)**. Chaque État membre doit en désigner au moins un. Ils sont chargés d'évaluer les demandes d'accès pour l'usage secondaire, de délivrer les permis et de superviser les environnements de traitement sécurisés. Ils garantissent le respect du RGPD. Les autorités numériques de santé nationales gèrent l'usage primaire au sein de l'infrastructure MyHealth@EU. Une coalition informelle de plus de 35 parties prenantes (EHTEL, etc.) apporte un rôle consultatif à la Commission européenne.

### Modèle de gouvernance

La gouvernance de l'EHDS est **hybride, décentralisée et fédérée**. Elle s'appuie sur un **Conseil EHDS** au niveau européen pour la coordination stratégique, co-présidé par la Commission. Au niveau national, les HDAB gèrent l'usage secondaire et les autorités sanitaires l'usage primaire. Les patients disposent d'un **droit d'opt-out** pour l'utilisation secondaire. Le modèle intègre des mécanismes de sanction en cas de non-conformité, avec des amendes comparables au RGPD, pour garantir la sécurité, la traçabilité et la confiance.

## Données échangées

Cette section décrit les données de santé électroniques échangées au sein de l'EHDS, en distinguant leur utilisation pour les soins (primaire) et pour la recherche/innovation (secondaire). Les échanges doivent garantir l'interopérabilité, la confidentialité et la qualité.

### Données d'usage primaire :

Ces données servent directement à la continuité des soins. Les patients peuvent y accéder et les partager avec des professionnels de santé dans un autre pays via des points d'accès nationaux. Les catégories prioritaires, à déployer progressivement d'ici mars 2029 et 2031, incluent :

* Les résumés de patient.
* Les ordonnances et dispenses électroniques (ePrescriptions/eDispensations).
* Les images médicales et leurs rapports.
* Les résultats de laboratoire.
* Les rapports de sortie d'hôpital.

### Données d'usage secondaire

Les données sont pseudonymisées et traitées dans des environnements sécurisés pour des finalités comme la recherche scientifique, l'entraînement d'IA et les politiques de santé publique. Les catégories couvrent les dossiers de santé électroniques, les données génomiques/-omiques humaines, les données de biobanques, les données d'applications bien-être, les données d'essais cliniques, les registres et les cohortes de recherche.

### Mécanismes d'échange

L'EHDS repose sur une architecture technique qui impose l'interopérabilité via des spécifications communes, des passerelles nationales et des API pour les systèmes de santé. Cette infrastructure s'appuie sur le RGPD, intégrant des contrôles de consentement et des options de retrait pour les patients, notamment dans le cadre des soins transfrontaliers.

**Cadre d'échange par usage**

* **Pour l'usage primaire (soins)** : L'accent est mis sur l'accès immédiat et la portabilité des données personnelles. Les échanges transfrontaliers sont facilités via l'infrastructure **MyHealth@EU**.
* **Pour l'usage secondaire (recherche/innovation)** : Les échanges sont hautement sécurisés et régulés. Ils s'effectuent via l'infrastructure **HealthData@EU**, où les données sont pseudonymisées et accessibles uniquement dans des environnements de traitement sécurisés (Secure Processing Environments - SPEs), sans possibilité de téléchargement des données brutes.

**Flux et gouvernance des accès**  
Le processus d'accès aux données secondaires est séquentiel et contrôlé :

1. Un demandeur (chercheur, industrie) soumet une requête à l'organisme national d'accès (HDAB).
2. Le HDAB évalue la demande et délivre un permis si les conditions sont remplies.
3. Les détenteurs de données (HDH) provisionnent les données pseudonymisées dans un SPE.
4. L'utilisateur accède et analyse les données au sein du SPE pendant une durée limitée.
5. Les résultats de l'analyse sont vérifiés avant sortie, et les données sont supprimées du SPE à l'expiration du permis.

Cette gouvernance stricte vise à équilibrer l'innovation avec la protection fondamentale de la vie privée et la souveraineté des données.

## Architecture du Data Space

L'architecture de l'EHDS est hybride, décentralisée et fédérée. Elle connecte les infrastructures nationales existantes sans centraliser les données, préservant ainsi la souveraineté des données au niveau national et institutionnel. Cette architecture s'articule autour de deux piliers opérationnels : **MyHealth@EU** pour l'usage primaire (soins) et **HealthData@EU** pour l'usage secondaire (recherche et innovation).

La structure repose sur plusieurs couches fonctionnelles :

* **Couche de stockage** : Basée sur une fédération décentralisée. Les données restent hébergées dans les systèmes nationaux ou institutionnels (silos de dossiers électroniques, systèmes hospitaliers). Il n'existe pas de base de données centrale européenne, ce qui garantit le contrôle local des données.
* **Couche de traitement** : Assure la préparation et la sécurisation des données. Elle comprend la standardisation des données vers des formats communs (comme FHIR via le format d'échange EEHRxF), ainsi que des processus obligatoires de pseudonymisation ou d'anonymisation pour l'usage secondaire. L'analyse des données secondaires s'effectue exclusivement dans des **Environnements de Traitement Sécurisés (SPE)** accrédités, empêchant le téléchargement des données brutes.
* **Couche d'exposition et d'accès** : Fournit l'interface sécurisée pour les utilisateurs via des API. L'accès est différencié :
  + Pour l'**usage primaire**, via **MyHealth@EU**, permettant aux professionnels de santé d'accéder aux dossiers patients en contexte transfrontalier.
  + Pour l'**usage secondaire**, via **HealthData@EU**, l'accès est strictement conditionné par un permis délivré par les organismes nationaux (HDAB) et limité aux finalités autorisées.
* **Services transversaux** : Des composants essentiels supportent l'ensemble de l'architecture :
  + **Catalogues de métadonnées** pour la découverte des jeux de données.
  + **Connecteurs standardisés** garantissant l'interopérabilité technique.
  + **Gestion des identités et des accès** pour l'authentification sécurisée.
  + **Sécurité** globale (chiffrement, audit) assurant la conformité avec le RGPD.

En synthèse, l'architecture de l'EHDS forme un écosystème interconnecté qui équilibre le partage de données à grande échelle avec la protection de la vie privée, la souveraineté des données et la confiance des citoyens.

## Schéma d’architecture

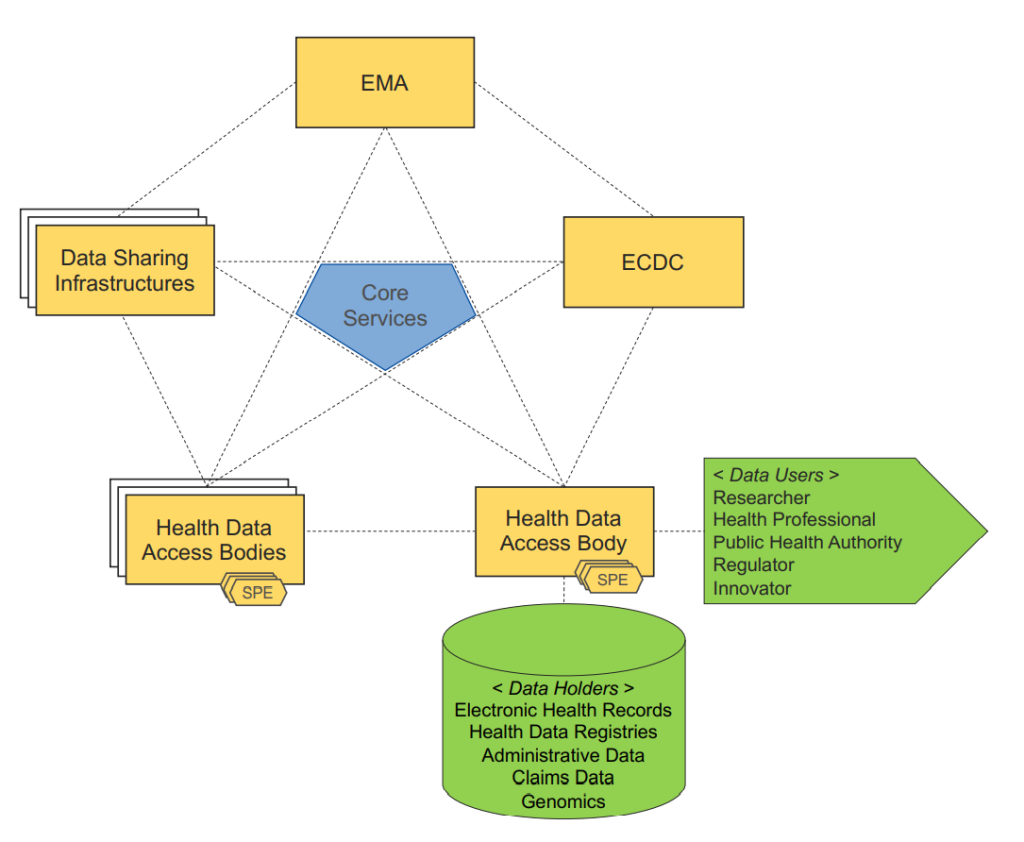


Figure 2 : L'architecture d'usage secondaire dans EHDS

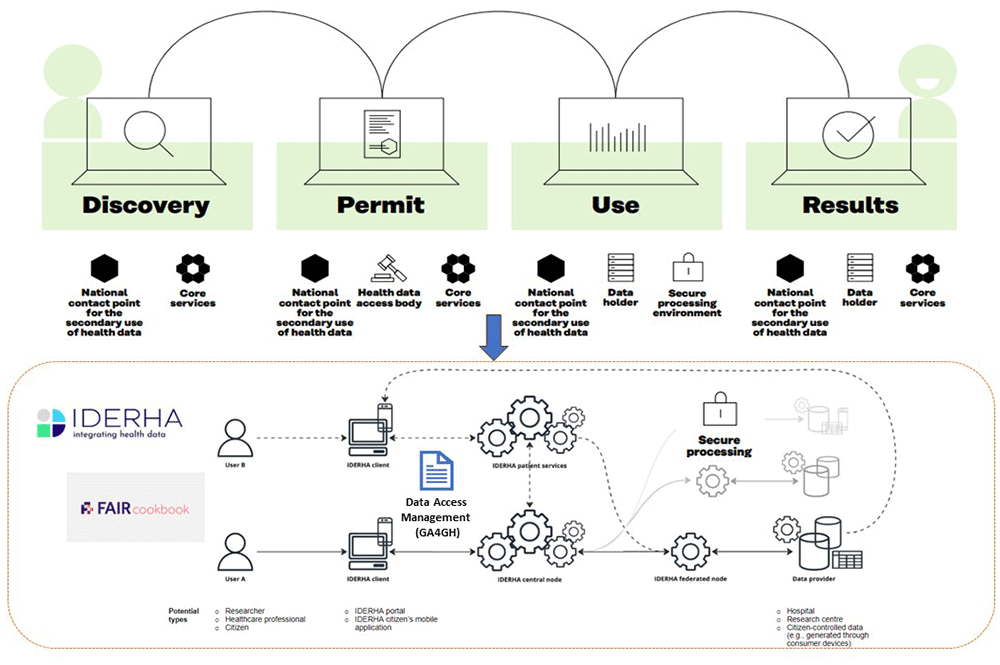


Figure 3 : L'infrastructure IDERHA en alignement avec le parcours utilisateur EHDS2

## Standards et technologies

L'EHDS s'appuie sur un ensemble cohérent de normes et de technologies pour garantir l'interopérabilité, la sécurité et la conformité au cadre réglementaire (RGPD, Règlement IA, NIS2). Ces choix sont alignés sur les principes FAIR et les recommandations de projets européens comme TEHDAS.

### Standards utilisés

Les standards se répartissent en trois catégories principales :

1. **Interopérabilité des systèmes de santé** :
   * **FHIR (Fast Healthcare Interoperability Resources)** : Standard pivot, basé sur des API, pour le format d'échange européen des dossiers de santé (EEHRxF). Il est obligatoire pour les systèmes de dossiers électroniques.
   * **HL7** : Fournit la base pour les guides d'implémentation européens, définissant des profils communs pour l'EHDS.
   * **Terminologies** : SNOMED CT (clinique), LOINC (laboratoire), ICD-11 (classification des maladies) pour l'harmonisation sémantique.
   * **DCAT-AP/HealthDCAT-AP** : Pour les catalogues de métadonnées facilitant la découverte des données.
2. **Web sémantique** :
   * **RDF/RDFS, OWL, SKOS** : Permettent de modéliser les données en graphes de connaissances, de créer des ontologies et de gérer des vocabulaires contrôlés pour une interopérabilité sémantique avancée.
   * **SPARQL** : Langage de requête pour interroger des données distribuées dans un format RDF, essentiel pour les environnements fédérés comme HealthData@EU.
3. **Sécurité et gestion des identités** :
   * **ISO/IEC 27001** : Norme pour la gestion de la sécurité de l'information, appliquée aux environnements de traitement sécurisés (SPE).
   * **eIDAS** : Règlement pour l'identification électronique et les services de confiance, garantissant un accès transfrontalier sécurisé.
   * **Pseudonymisation/Anonymisation** : Processus obligatoires pour l'usage secondaire, conformes au RGPD.
   * **Apprentissage fédéré** : Permet l'analyse des données sans les déplacer, préservant la souveraineté.

### Justification des choix

Ces standards sont sélectionnés pour leur contribution spécifique aux objectifs de l'EHDS :

* **Interopérabilité sémantique** : FHIR, combiné aux technologies du Web sémantique (RDF, OWL), harmonise les données hétérogènes sans perte de sens, facilitant la recherche et l'innovation (ex. : médecine personnalisée).
* **Échanges transfrontaliers** : Les API standardisées (FHIR) et la gestion d'identité (eIDAS) permettent une continuité des soins fluide et une recherche collaborative à l'échelle de l'UE, réduisant les silos nationaux.
* **Sécurité des données** : Les normes comme ISO 27001 et les processus de pseudonymisation intégrés dans les SPE protègent la confidentialité et préviennent les risques de réidentification, renforçant la confiance publique.
* **Évolutivité** : La modularité de FHIR et l'approche fédérée (apprentissage distribué, requêtes SPARQL) assurent une architecture scalable, capable d'intégrer de nouveaux types de données (génomiques) et cas d'usage (IA) pour les 27 États membres.

Ces choix techniques soutiennent directement les objectifs stratégiques d'empowerment des patients, d'innovation responsable et de gouvernance décentralisée de l'EHDS.

# PARTIE II : Intégration des données

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquementL’intégration des données constitue un pilier central de la mise en œuvre de l’Espace européen des données de santé (EHDS). Elle vise à rendre des données de santé exploitables de manière **sécurisée**, **interopérable** et **réutilisable**, malgré l’hétérogénéité des systèmes de soins, des formats, des modèles et des terminologies.  
Cette partie suit strictement les exigences du cahier des charges : **analyse des besoins**, **description des sources (≥ 3)**, **architecture d’intégration**, **flux d’intégration**, puis **prototype expérimental** (avec preuves : captures, scripts, extraits).

Figure 3 Arborescence du projet et des dataset intégrés

## Analyse des besoins d’intégrations

Dans le cadre du EHDS, l’intégration des données répond à deux catégories de besoins distinctes mais complémentaires : **l’usage primaire** (soins) et **l’usage secondaire** (recherche, innovation, politiques publiques).

* **Usage primaire (soins)** : nécessite des échanges fréquents voire quasi temps réel, notamment en contexte transfrontalier. L’intégration s’appuie sur des **API sécurisées** et des **standards d’interopérabilité** (ex. FHIR), afin de garantir l’exactitude, la cohérence et la mise à jour des données cliniques.
* **Usage secondaire (recherche)** : concerne l’exploitation de volumes importants (données historiques et multi-sources). Les traitements sont majoritairement **batch** et impliquent des exigences renforcées : **pseudonymisation**, **traçabilité**, **qualité des données** et **gouvernance des accès**. Les analyses doivent être réalisées dans des environnements de traitement sécurisés (type **SPE**) sans export des données brutes.

Les principaux problèmes identifiés sont :

* **Silos** nationaux/institutionnels et difficulté de consolidation ;
* **Hétérogénéité** formats / modèles / terminologies ;
* **Incohérences** et lacunes de qualité ;
* **Délais** d’accès et de traitement incompatibles avec certains usages.

L’intégration vise ainsi à dépasser ces limites, conformément aux principes **FAIR**.

Une image contenant texte, capture d’écran, Police, Page web

Description générée automatiquement

Figure 4 Besoin d'intégration EHDS

## Source de données retenues pour le prototype

Afin de répondre aux exigences du projet et de disposer d’un cadre expérimental réaliste et reproductible, nous avons intégré **plus de trois sources hétérogènes** :

1. **EHR CSV (source\_ehr\_csv/ehr\_patients.csv)**  
   Représente une extraction “classique” de dossier patient (format tabulaire), utile pour tester l’intégration structurée.
2. **FHIR NDJSON (source\_fhir\_ndjson/bundle.ndjson)**  
   Données cliniques structurées au format FHIR-like (patient, condition, prescription, allergy). Elles simulent l’interopérabilité attendue par EHDS pour l’usage primaire.
3. **Laboratoire JSON (source\_lab\_json/lab\_results.json)**  
   Données de biologie médicale (hétérogénéité d’unités, valeurs manquantes, standardisation) : cas typique d’intégration multi-sources.
4. **Imagerie DICOM (source\_dicom/...)**  
   Données d’imagerie au format DICOM (métadonnées + fichiers binaires). Elles illustrent les contraintes spécifiques : volume, structure complexe, normalisation et gouvernance.
5. **MIMIC-III (mimic/...)**  
   Jeu de données clinique (usage secondaire, batch, tables volumineuses), pertinent pour simuler les données de recherche.

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Figure 5 Preuve de FHIR

Une image contenant texte, logiciel, Icône d’ordinateur, Page web

Description générée automatiquement

Figure 6 Preuve DICOM et lab\_results.json

**Tableau récapitulatif des sources**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | Format | Volume Estimé | Fréquence de Mise à Jour | Usage | Contraintes |
| Synthea (Dossiers Médicaux Électroniques) | FHIR (JSON/XML) | 1-10 Go (selon nombre de patients simulés) | Quasi-temps réel pour usage primaire | Primaire (soins patients) | RGPD (pseudonymisation), FAIR, hétérogénéité sémantique |
| MIMIC-III (Données Cliniques) | CSV | ~2 Go | Batch (historique) | Secondaire (recherche) | Anonymisation renforcée, traçabilité, qualité des données |
| Données de Laboratoire Simulées | CSV/JSON | 100 Mo - 1 Go | Périodique (quotidienne/hebdomadaire) | Primaire/Secondaire | Harmonisation unités, gestion valeurs manquantes, RGPD |
| Imagerie Médicale Synthétique (Synthea Coherent Data Set) | DICOM | 5-50 Go | Batch (ponctuelle) | Primaire/Secondaire | Normalisation formats binaires, volumes élevés, alignement sémantique, RGPD |

## Architecture d’intégration des données

### Comparaison des architectures

Pour concevoir l'architecture d'intégration dans l'EHDS, quatre approches modernes sont comparées : Data Lake, Data Warehouse, Data Mesh et Data Fabric. Le tableau ci-dessous résume leurs caractéristiques, avantages et inconvénients dans le contexte EHDS (fédéré, souveraineté des données, RGPD, usages primaire/secondaire).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Architecture | Centralisation | Gouvernance | Interopérabilité | Souveraineté des données | Adaptée à EHDS ? |
| Data Lake | Centralisée | Faible (schema on read) | Moyenne | Faible (duplication) | Non : risque RGPD, pas fédérée |
| Data Warehouse | Centralisée | Forte (schema on write) | Bonne (structurée) | Faible | Non : coûteux, pas scalable pour volumes santé |
| Data Mesh | Décentralisée | Domaines autonomes | Haute (produits data) | Haute | Oui : alignée fédération, domaines (hôpitaux) |
| Data Fabric | Hybride/fédérée | Automatisée (metadata) | Très haute | Haute | Oui : unifie sans déplacer données |

EHDS favorise une architecture fédérée. C’est pour cela Data Fabric est le choix adopté selon l’organisation.

**Data Lake (stockage brut centralisé)**

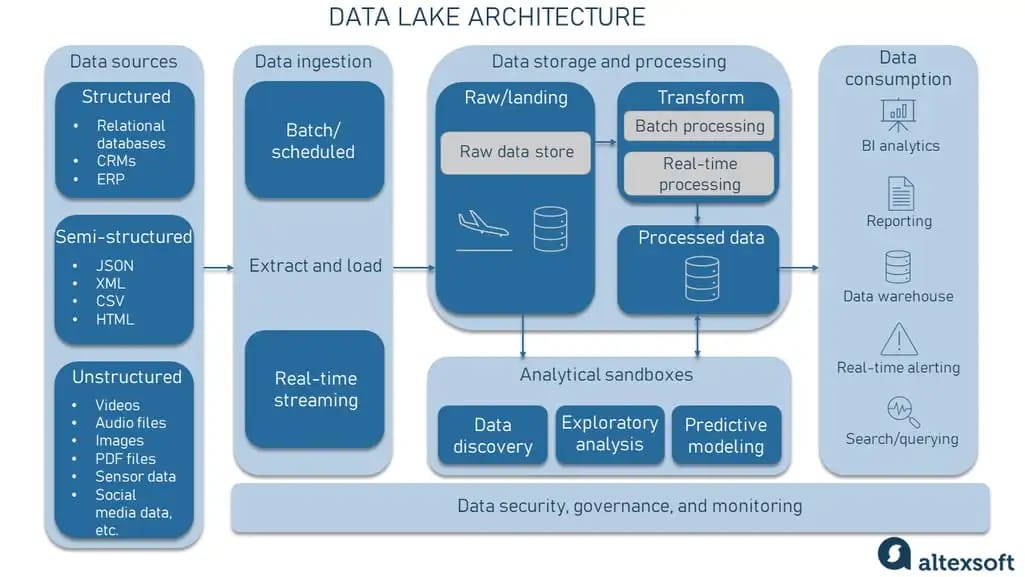


Figure 7 Architecture Data Lake

**Data Warehouse (Données structurées centralisé)**

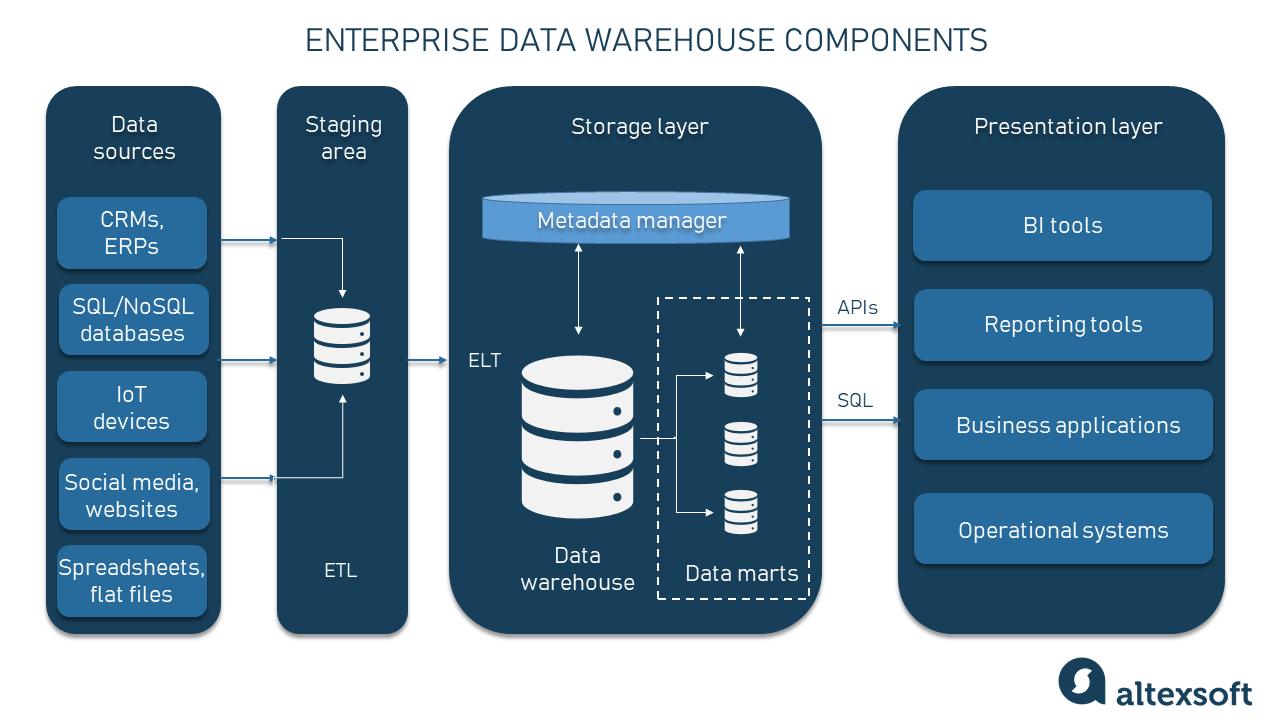


Figure 8 Architecture Data Warehouse

**Data Mesh (Domaines décentralisés)**

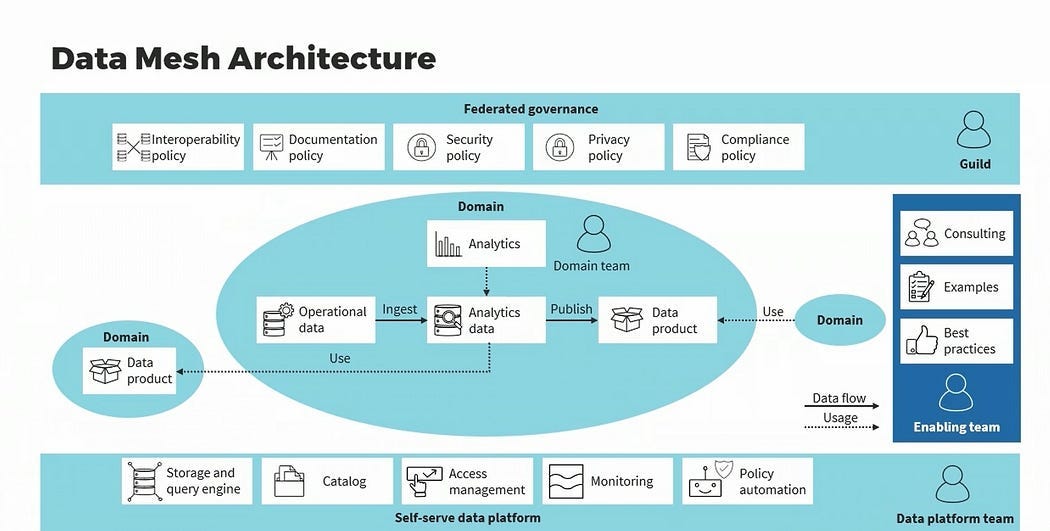


Figure 9 Architecture Data Mesh

### Comparaison des architectures

L’architecture adoptée est une **architecture fédérée hybride**, inspirée des principes du **Data Fabric** et intégrant des éléments du **Data Mesh**, en cohérence avec la vision EHDS.

* **Pourquoi pas Data Lake / Data Warehouse ?**  
  Ces architectures reposent sur une centralisation massive des données, ce qui augmente les risques en santé (confidentialité, RGPD, souveraineté, duplication).
* **Pourquoi pas Data Mesh seul ?**  
  Le Data Mesh est décentralisé mais nécessite une couche de médiation et de gouvernance pour garantir harmonisation et interopérabilité.

Dans l’architecture retenue :

* Les données restent hébergées chez les détenteurs (local data holders).
* Une couche de médiation (pipeline ETL open-source) assure extraction, harmonisation, pseudonymisation et interopérabilité.
* L’accès est contrôlé via une logique de traitement sécurisé (SPE simulé), alignée avec HealthData@EU.

Une image contenant texte, capture d’écran, Police, Page web

Description générée automatiquement

Figure 10 Architecture d'intégration fédérée proposée

## Flux d’intégration des données

Les flux d’intégration sont organisés selon la finalité de l’usage. Les traitements batch sont privilégiés pour l’intégration des jeux de données historiques et volumineux, tandis que des flux quasi temps réel peuvent être envisagés pour les données liées à l’usage primaire.

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure 11 Flux d'intégration (batch vs quasi temps réel)

Le processus d’intégration comprend les étapes suivantes :

1. Extraction des données depuis les sources hétérogènes.
2. Transformation et normalisation vers des formats et standards communs.
3. Nettoyage et amélioration de la qualité des données.
4. Pseudonymisation des identifiants patients.
5. Chargement des données dans un espace intégré sécurisé.

Pour les usages secondaires, l’ensemble des traitements est exécuté dans un environnement simulant un **Secure Processing Environment (SPE)**, conformément aux exigences du EHDS, interdisant toute extraction directe des données brutes.

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Figure 12 La pseudonymisation (patient\_id)

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Figure 13 La normalisation unité Labo

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Figure 14 Le parsing FHIR

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement

Figure 15 La parsing DICOM (métadonnées)

## Prototype d’intégration : description, outils , résultats

Le prototype implémenté vise à démontrer expérimentalement l’intégration de plusieurs sources hétérogènes de santé, conformément aux principes EHDS. Le pipeline produit :

* Une base intégrée ehds.db (tables patients, labs, fhir, dicom, etc.)
* Un graphe RDF ehds\_data.ttl
* Une visualisation via Streamlit dashboard.py

**Outils utilisés**

* Python (ETL)
* SQLite (stockage intégré)
* rdflib (RDF + SPARQL)
* Streamlit + Plotly (dashboard)
* pydicom (DICOM)

**Résultats obtenus**

* Le pipeline génère et intègre automatiquement les sources.
* Les tables intégrées permettent des jointures patient↔lab↔FHIR↔imagerie.
* Le graphe RDF rend possible des requêtes sémantiques (Partie III).

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Figure 16 Terminal Exécution complète du pipeline

Une image contenant texte, capture d’écran, logiciel, affichage

Description générée automatiquement

Figure 17 Base intégrée SQLite

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

Figure 18 Extrait du graphe RDF

Une image contenant texte, logiciel, Icône d’ordinateur, Page web

Description générée automatiquement

Figure 19 Vue global Dashboard

# PARTIE III : Interopérabilité sémantique

# Conclusion

# Références

## ****Références réglementaires et institutionnelles****

1. Commission européenne. Proposal for a Regulation on the European Health Data Space (EHDS), 2022.
2. Union européenne. Règlement (UE) 2025/327 relatif à l’Espace européen des données de santé, Journal officiel de l’Union européenne, 5 mars 2025.
3. European Data Protection Board (EDPB). Opinions and Guidelines on the European Health Data Space and GDPR, 2022–2024.
4. TEHDAS2 Joint Action. Governance and implementation reports on the European Health Data Space, 2023–2025.
5. EHTEL. European Health Data Space: Strategic initiative or a document with weaknesses?, 26 février 2024.

## ****Rapports, études et documents de référence****

1. EIT Health Think Tank. Implementing the European Health Data Space across Europe, avril 2024.
2. European Commission. European Health Data Space – the first of many, présentation institutionnelle.
3. de Barros, A. The European Health Data Space: secondary use of health data, DG SANTE C1.
4. SIEPS. European Policy Analysis on the European Health Data Space, février 2024.
5. European Parliamentary Research Service (EPRS). European Health Data Space, Briefing, 2024.
6. European Health Data Space: An Opportunity Now to Get It Right, Healthcare Policy Article, 2022.

## ****Sources pédagogiques et supports de cours****

1. Chapitre 1 – Data Spaces, support de cours, 2025.
2. Chapitre 2 – Intégration des données, support de cours, 2025.
3. Projet 2025 – Module AISD, document pédagogique.

## ****Sources web et veille documentaire****

1. <https://healthdataspace.eu>
2. <https://www.european-health-data-space.com>
3. <https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_en>
4. <https://health.ec.europa.eu/ehealth-digital-health-and-care/european-health-data-space-regulation-ehds_fr>
5. <https://eur-lex.europa.eu/eli/reg/2025/327/oj>
6. <https://esante.gouv.fr/espace-europeen-donnees-sante>
7. <https://www.ihe-europe.net/european-health-data-space>
8. <https://eithealth.eu/wp-content/uploads/2024/04/EIT_Health_ThinkTank_Implementing_the_EHDS_across_Europe_23.04.24.pdf>
9. <https://pmc.ncbi.nlm.nih.gov/articles/PMC12420905/>
10. <https://cromospharma.com/the-european-health-data-space-ehds-a-game-changer/>
11. <https://www.has-sante.fr>
12. <https://www.health-data-hub.fr>