

Spécifications techniques et fonctionnelles

Standardisation et interopérabilité du socle commun

Introduction	1
1. Rappel du contexte du groupe de travail	1
2. Consignes de lecture et mode d'emploi	2
3. Références et documentation utiles	2
Éléments de standardisation des données d'identification du socle commun	2
1. Extrait des items du socle	2
2. Standardisation des items	4

Introduction

1. Rappel du contexte du groupe de travail

Dans la continuité du groupe de travail (GT) "Ensemble commun de données", qui a établi en 2023 un socle commun aux entrepôts de données hospitaliers (EDSH), le comité stratégique des données de santé a mandaté début 2024 un nouveau groupe de travail pour proposer des standards pour les 51 items qui constituent le socle. Cette étape est essentielle pour permettre son utilisation harmonisée et sa mutualisation.

Les représentants des EDS hospitaliers lauréats de l'appel à projet (AAP) France 2030 soutenant le déploiement des EDSH et d'institutions nationales liées aux données de santé et à l'interopérabilité se sont penchés sur les cinq familles de variables du socle (données démographiques, données du PMSI, données biologiques, données du médicament et données d'examens cliniques) pour identifier les éléments de standardisation les plus pertinents au sein du modèle OMOP et consigner ces choix, arbitrages et préconisations au sein de livrables.

Pour plus d'informations sur le contexte du groupe de travail, consultez la note de synthèse des travaux.

2. Consignes de lecture et mode d'emploi

Méthodologie du groupe de travail "Standardisation et interopérabilité du socle commun aux EDSH"

Le groupe de travail s'est organisé en sous-groupes de réflexion, répartis sur chacune des cinq thématiques du socle commun (identité, biologie, médicaments, PMSI, données cliniques et style de vie).

Chaque sous-groupe s'est d'abord penché sur la standardisation de son ensemble de données, puis les 5 sous-groupes ont restitué leurs propositions à l'ensemble du GT pour validation et pour s'accorder sur les préconisations adéquates compte-tenu des contraintes réelles des EDSH.

Ce livrable final a pour objectif de restituer les choix préconisés par le GT, afin de permettre

aux entrepôts de données de santé de mettre en oeuvre le socle commun au sein de leur structure et de le partager de manière standardisée

Il s'accompagne également d'un fichier Excel avec les tables OMOP utilisées pour la standardisation et les variables associées.

Travaux et décisions du GT

Le groupe de travail a choisi le standard OMOP-CDM comme référence afin de spécifier les données du socle commun.

En effet, OMOP est maîtrisé par la majorité des membres du GT et permet de répondre aux objectifs du socle commun :

- Créer des bases standardisées pour réaliser des études communes sur le socle de données,
- Pouvoir interroger toutes les données de la même façon,
- Pouvoir interpréter les résultats de la même manière, peu importe leur EDS d'origine, grâce à une sémantique commune et représentation homogène des données.

En complément des travaux sur le modèle OMOP, un sous-groupe a été ajouté pour réfléchir à une manière homogène et opérationnelle de documenter les solutions retenues au format FHIR. Ces travaux étant encore au stade expérimental, ils n'ont pas été intégrés dans ces spécifications, mais sont mis à disposition sur un Github public : [Lien vers le répertoire Github contenant les travaux FHIR.](#)

Notes des rédacteurs et rédactrices

Les livrables du GT "Standardisation et interopérabilité du socle commun aux EDSH" contiennent les choix du groupe de travail quant à la standardisation des données du socle commun aux EDSH uniquement. Ils n'ont pas vocation à couvrir l'ensemble des données des EDSH et pourront évoluer avec le temps. Les porteurs d'EDS restent libres des choix de standard au sein de leurs structures.

Pour toute question ou remarques sur ces documents veuillez vous adresser à l'adresse suivante : contact@health-data-hub.fr.

3. Références et documentation utiles

- Socle commun défini dans le cadre du groupe de travail « Ensemble des variables communes à tous les entrepôts de données de santé »
- Note de synthèse du groupe de travail « standards et interopérabilité »
- Ressources OMOP-CDM :
 - Standardized Data: The OMOP Common Data Model - OHDSI
 - OMOP Common Data Model - Github
 - Athena – OHDSI Vocabularies Repository

Éléments de standardisation des données d'identification du socle commun

1. Extrait des items du socle

Groupe d'items	Intérêt du recueil	Conditions	Item	Définition de l'item	Valeur attendue	Métadonnées
1.a. Identité patient	Identitologie Clé d'appariement unique pouvant permettre le chaînage direct du socle avec d'autres bases	Au moins un identifiant attendu	Numéro d'inscription au Répertoire (NIR)	Numéro unique attribué à chaque personne à sa naissance sur la base d'éléments d'état civil transmis par les mairies à l'INSEE	Chaîne de caractères	N/A
			Identité Nationale de Santé (INS)	Numéro d'identité unique, pérenne, partagée par l'ensemble des professionnels du monde de la santé	Chaîne de caractères	N/A
	Identitologie		Date de naissance	Date de naissance des papiers d'identité utilisés pour la production de l'INS	Date	N/A
			Date de décès	Date de décès à l'hôpital, ou date de décès collectée par chaînage avec une base externe comme l'INSEE ou le Cépidec	Date	Source de la donnée

			Rang gémellaire du bénéficiaire	Pour le régime général, il permet de distinguer les naissances gémellaires de même sexe	Chaine de caractères	N/A
1.b. Environnement	Clé d'appariement avec les bases de données environnementales		Adresse géocodée	Coordonnées géographiques (latitude et longitude) de l'adresse du patient	Coordonnées géographiques	Date du recueil de l'information
	Données essentielles pour la conduite de recherches et d'études liées aux déterminants de santé	Si adresse géocodée (1.b.1) non disponible	IRIS de l'adresse	Codage de l'adresse du patient selon la méthode d'Ilots Regroupés pour l'Information Statistique de l'INSEE	Chaîne de caractères	Date du recueil de l'information

2. Standardisation des items

a. Recommandations générales de liaison des informations au séjour

Le GT recommande de **relier les différents items du socle à la notion de "visit"** de la table "visit_occurrence", comprise comme un contact pour un patient avec un établissement de santé :

- Une visit_occurrence commence lors de la venue du patient dans l'établissement de santé (cela exclut donc la planification de sa venue) pour une modalité de prise en charge spécifique (consultation, acte externe, hospitalisation complète, passage aux urgences...).
- Une visit_occurrence se termine lors de la sortie du patient de l'établissement de santé pour la même modalité de prise en charge.

La définition d'établissement de santé reste à la discrétion de chaque structure (FINESS juridique ou géographique).

Ces modalités de définition de la notion de "visit" induisent les conséquences suivantes :

- Plusieurs résumés PMSI peuvent donner une seule visit_occurrence (Résumé d'Unité Médicale, Résumés Hebdomadaires Standardisés, etc)
- Une même venue avec des modalités de prise en charge différentes produisent plusieurs visit_occurrence. Par exemple, un patient venant en consultation puis hospitalisé à la suite de la consultation entraînera la création de 2 "visit_occurrence".

Pour les séjours, le GT recommande aux établissements d'alimenter également la table "visit_detail" afin de **repérer les unités médicales particulières traversées par les patients** (par exemple, réanimation ou psychiatrie), mais **cette alimentation resterait à la discrétion des établissements**, ainsi que la mise en lien avec les tables regroupant les actes, les diagnostics, etc. qui serait laissée à l'initiative des établissements.

La description de la standardisation des items liés au séjour est disponible dans le livrable 02A_GT Standardisation - spécifications - données PMSI.

b. Items "Numéro d'inscription au Répertoire (NIR)", "Identité Nationale de Santé (INS)", et "rang gemellaire du bénéficiaire"

i. Description

NIR (Numéro d'Inscription au Répertoire) : C'est un identifiant unique attribué à chaque personne née en France. Il est composé de 13 chiffres qui indiquent le sexe, l'année de naissance, le mois de naissance et le lieu de naissance de l'individu, suivi d'un numéro d'ordre. Le NIR est également connu comme le numéro de sécurité sociale (NSS).

NIA (Numéro d'Immatriculation d'Attente) : Il est attribué à toute personne qui ne dispose pas d'un NIR mais qui remplit les conditions d'affiliation à un régime obligatoire de sécurité sociale. Le NIA est constitué sur le même modèle que le NIR, il est transformé en NIR dès que l'on a vérifié qu'il n'existait aucun NIR identique déjà attribué.

Matricule INS (Identité Nationale de Santé) : C'est un numéro d'identité unique, pérenne, partagé par l'ensemble des professionnels du monde de la santé. Le matricule INS est basé sur le NIR ou le NIA de l'usager. Le référentiel INS et le RNIV stipulent que le matricule INS

doit être associé systématiquement à une information précisant sa nature : NIR ou NIA. Pour le régime général, le **rang gémellaire du bénéficiaire** permet de distinguer les naissances gémellaires de même sexe.

ii. Stratégie de standardisation OMOP

Le matricule INS et le NSS (numéro de sécurité sociale) sont tous les deux basés sur le NIR ou le NIA de l'individu. Toutefois, une même personne peut avoir plusieurs NIR ou NIA au cours de sa vie : passage à l'âge adulte, changement de sexe, enregistrement administratif, etc. Il paraît important de pouvoir retracer l'historique de ces changements pour un même individu.

L'utilisation des tables du modèle OMOP ne permet pas de stocker l'historique des NIR et des INS. Par ailleurs, dans ATHENA, seul existe un concept SNOMED CT pour le numéro de sécurité sociale, qui ne permet donc pas de différencier le NIR de l'INS.

Pour ces raisons, le groupe de travail propose la création d'une table spécifique pour stocker ces informations.

Cette table, nommée PERSON MATCHING ATTRIBUTE, contiendrait les champs suivants :

- person_matching_attribute_id : clé primaire identifiant unique de l'enregistrement
- person_id : clé étrangère vers la table PERSON
- person_identifier_concept_id : clé étrangère vers la table CONCEPT permettant de décrire le type d'attribut
 - Valeurs possibles locales : "NIR", "INS-NIR", "INS-NIA", "rang gémellaire"
 - Exemple : "NIR"
- value_as_string : valeur de l'attribut, pour le NIR ou l'INS
 - Exemple : "2011275116056"
- value_as_number : valeur de l'attribut, pour le rang gémellaire
 - Exemple : "2"

Champ	Obligatoire	Format	Clé primaire	Clé étrangère	Fk Table	Fk Domaine
person_matching_attribute_id	Oui	INTEGER	Oui	Non		
person_id	Oui	INTEGER	Non	Oui	PERSON	
person_identifier_concept_id	Oui	INTEGER	Non	Oui	CONCEPT	OBSERVATION
value_as_string	Oui	VARCHAR(15)	Non	Non		
value_as_number	Oui	INTEGER	Non	Non		

La standardisation de ces items ne fait pas appel à des terminologies standard nationales ou internationales. Ces items s'appuient sur des concepts locaux qui devront être communs à tous les EDS.

c. Item "Date de naissance"

i. Description

Date de naissance des papiers d'identité utilisés pour la production de l'INS.

ii. Stratégie de standardisation OMOP

Table OMOP utilisée : Person

Champs utilisés :

- **"year_of_birth"** : la seule variable concernant la date de naissance obligatoire dans OMOP
- **"month_of_birth"** : renseigner avec le mois de naissance, laisser vide si inconnue
- **"day_of_birth"** : renseigner avec le jour de naissance, laisser vide si inconnue
- **"birth_datetime"** : la préconisation du GT est de toujours renseigner cette variable même si elle n'est pas obligatoire avec la date et l'heure de naissance si elle est connue car elle peut être utile notamment pour les études de néonatalogie

Dans le cas d'intégration de données d'études, pour lesquelles seule une partie de la date de naissance a été recueillie, le GT préconise de suivre les recommandations OMOP pour le remplissage du champ birth_datetime :

- Si le jour de naissance est vide et le mois de naissance n'est pas vide, il convient d'utiliser le premier jour du mois et de l'année de naissance
- Si le mois de naissance est vide ou si le jour de naissance et le mois de naissance sont tous les deux vides et que le patient a des enregistrements au cours de son année de naissance, utilisez la date de l'enregistrement le plus ancien, sinon utilisez le 15 juin de l'année de naissance
- Si l'heure de naissance n'est pas indiquée, utiliser minuit (00:00:0000).

Cas d'une date imprécise : lorsqu'un patient ne connaît pas sa date de naissance exacte ou lorsque la date de naissance fournie par le document d'identité où le dispositif d'identification numérique est incomplète, une date approximative est saisie dans le logiciel de gestion administrative des patients selon les règles du Référentiel National d'Identitovigilance (RNIV).

Le caractère imprécis de la date de naissance n'est pas toujours indiqué et les règles de substitution des informations absentes ont évoluées au fil du temps.

Le GT préconise dans ce cas de remplir les champs avec la date de naissance disponible et de renseigner une observation pour documenter le caractère inconnu (concept_id=40318436 « Patient DOB unknown ») ou imprécis de la date de naissance (concept_id=40318453 « Questionable if patient date of birth correct »).

Table	Champ	Format
person	year_of_birth	integer
person	month_of_birth	integer

person	day_of_birth	integer
person	birth_datetime	datetime

d. Item "Date de décès" et donnée associée "source de la donnée"

i. Description

Date de décès : Date de décès à l'hôpital, ou date de décès collectée par chaînage avec une base externe comme l'INSEE ou le CépiDc.

Source de la donnée : Métadonnée associée à l'item "date de décès"

ii. Stratégie de standardisation OMOP

Table OMOP utilisée : Death

Champs utilisés :

- **"death_date"** : la seule variable concernant la date de décès obligatoire dans OMOP
- **"death_datetime"** : la préconisation du GT est de toujours renseigner cette variable même si elle n'est pas obligatoire avec la date et l'heure de décès si elle est connue. Si l'heure de décès n'est pas indiquée, utiliser minuit (00:00:0000).
- **"death_type_concept_id"** : ce champ doit permettre d'indiquer la source de la date de décès. Elle fait référence à des concepts standards OMOP :
 - o Si l'information provient d'un outil du DPI de l'établissement, utiliser Concept_id=32817 "EHR",
 - o Si l'information provient d'une source externe vérifiée (CépiDC, intégration dans la GAM du certificat de décès), utiliser Concept_id=32815 " Death Certificate
 - o Si l'information provient d'un appariement avec le fichier public des décès publié par l'INSEE (Fichiers des personnes décédées depuis 1970 | Insee), utiliser Concept_id=32848 : "Government report"

Table	Champ	Format
death	death_date	date
death	death_datetime	date time
death	death_type_concept_id	integer

Remarque : le GT souligne qu'il n'y a aujourd'hui pas de méthode de rapprochement probabiliste homogène entre les données des EDS et les données de décès issues du fichier des personnes décédées depuis 1970 de l'INSEE.

Il paraît intéressant que les acteurs documentent la performance de leur algorithme d'appariement.

d. Item « Adresse géocodée » et donnée associée « date de recueil de l'information »

i. Description

Adresse géocodée : Coordonnées géographiques (latitude et longitude) de l'adresse du patient.

Date de recueil de l'information : Date de recueil du géocodage

ii. Standardisation OMOP

Table OMOP utilisée : location

Champs OMOP utilisés :

Champ	ETL Conventions	Format
latitude	Must be between -90 and 90	float
longitude	Must be between -180 and 180	float

• Précisions :

La latitude et longitude seront au format WGS84, degré décimal.

La qualité du géocodage dépend de la qualité de l'adresse recueillie. Si aucune adresse n'est disponible pour le patient ou de trop mauvaise qualité pour réaliser le géocodage, les champs pourront être laissés vides. Le groupe de travail préconise d'indiquer le meilleur géocodage possible.

Le groupe de travail recommande de documenter la méthode de géocodage utilisée (correspondance stricte, ou fuzzy finding ; package opensource, par exemple Addok, ou développement local) et de décrire le niveau de qualité atteint.

Concernant la date de recueil de l'information, ni la table LOCATION, ni la table PERSON ne permettent de stocker une date de recueil de l'information. Les adresses sont mises à jour lors de venues dans l'établissement. Par convention, on considérera donc que la date de recueil du géocodage est la date la plus récente du champ visit_start_date de la table VISIT_OCCURENCE pour le patient. Le groupe de travail préconise de mettre à jour les données de géocodage lors de chaque visite du patient dans l'établissement de santé.

e. Item "IRIS" et donnée associée "date de recueil de l'information"

i. Description

IRIS : Codage de l'adresse du patient selon la méthode d'Ilots Regroupés pour l'Information

Statistique de l'INSEE.

ii. Stratégie de standardisation OMOP

Table OMOP utilisée : location

Champs OMOP utilisés :

Champ	Format
county	varchar(20)

- **Précisions :**

Le groupe de travail recommande de documenter la méthode utilisée pour déterminer l'IRIS et la version de l'IRIS.

Concernant la date de recueil de l'information, ni la table LOCATION, ni la table PERSON ne permettent de stocker une date de recueil de l'information. Les adresses sont mises à jour lors de venues dans l'établissement. Par convention, on considérera donc que la date de recueil de l'IRIS est la date la plus récente du champ visit_start_date de la table VISIT_OCCURENCE pour le patient. Le groupe de travail préconise de mettre à jour les données de l'IRIS lors de chaque visite du patient dans l'établissement de santé.