AGENCE
DU NUMÉRIQUE
EN SANTÉ

Constitution d'un corpus sémantique national : Faut-il adopter la SNOMED CT ?

Rapport final – Octobre 2020

Classification : Publique *Version : Finale*

Agence du Numérique en Santé
9, rue Georges Pitard – 75015 Paris
T. 01 58 45 32 50
esante.gouv.fr

SOMMAIRE

TABLES DES ILLUSTRATIONS

FIGURES

TABLEAUX

Note de lecture

Le secteur santé-social est associé à de nombreux domaines de connaissances décrits par des ressources sémantiques variées.

Ces ressources sémantiques constituent le langage commun et partagé par tous les experts santé-social. Elles représentent ainsi le **socle nécessaire** au développement de la digitalisation des échanges de santé en production de soins, pilotage d'activité, ou exploitation de données en recherche.

Selon leur complexité, plusieurs types de ressources peuvent être distingués :

- liste / dictionnaire / glossaire ;
- taxonomie / classification ;
- thésaurus ;
- terminologie ;
- ontologie.

Ces différentes ressources sont précisément décrites dans une étude bibliographique (**annexe P3.0**).

Dans ce rapport, de manière générique, les termes ressources sémantiques ou Terminologie seront employés pour désigner indifféremment dictionnaire, thésaurus ou ontologies.

A noter que le terme ressource termino-ontologique (RTO) ou système d'organisation de la connaissance (SOC) sont utilisés à certains passages du rapport pour désigner des terminologies ou des ontologies de manière indifférenciée[1].

[1] https://www.irit.fr/publis/MELODI/AussenacCharletReynaudDelaitre_book-cepadues-chap-IC-2014.pdf

1 INTRODUCTION

1.1 La gouvernance des Terminologies en France : le CGTS

Missionnée par la Délégation ministérielle du Numérique en Santé (DNS, ex-DSSIS), l'Agence du Numérique en Santé (ANS, ex-ASIP Santé) a mené entre 2014 et 2017 une série de travaux sur les terminologies du domaine santé et social. L'objectif était alors d'étudier le choix et l'adoption éventuelle d'un référentiel d'interopérabilité sémantique pour le domaine. La quatrième phase de cette étude portait sur les conditions de mise en œuvre de terminologies de référence. Elle visait à instruire l'opportunité d'acquérir la SNOMED CT et à préciser la trajectoire d'une éventuelle adhésion à SNOMED International.

Ces travaux ont conclu :

- Que les conditions requises (gouvernance notamment et identifications claires des cas d'usages) n'étaient pas réunies pour une adoption immédiate ;
- Qu'il est nécessaire d'évaluer des terminologies candidates à la position de référence avant adoption et généralisation ;
- Qu'il est prioritaire de mettre en place une maîtrise d'ouvrage opérationnelle de mise à disposition des terminologies de santé.

Le Comité de Pilotage d'avril 2018 a jugé, en accord avec la DNS, que cette maîtrise d'ouvrage devait être confiée à l'ANS. C'est le pôle des affaires médicales et labellisation (PML) de l'ANS qui a pris en charge, dès 2018, les travaux de mise en œuvre du Centre de Gestion des Terminologies de Santé (CGTS). Le CGTS répond aux enjeux :

- de besoins croissants en ressources sémantiques face à l'augmentation des échanges dématérialisés dans le secteur santé-social ;
- de besoins de guichet national de publication de ressources sémantiques exprimés par les parties prenantes de l'écosystème du numérique en santé ;
- de facilitation de l'accès aux ressources sémantiques tout en garantissant leur exploitabilité.

Le CGTS a 3 missions : maîtrise d'ouvrage opérationnelle des ressources sémantiques, guichet national de publication et accompagnement des utilisateurs. Par ailleurs, le centre entend entreprendre le travail de rationalisation de l'offre des terminologies médicales, afin de constituer un corpus national cohérent, composé des ressources nécessaires à l'écosystème, en commençant par organiser l'effort qualitatif qui doit être mené autour des terminologies déjà en usage.

Le CGTS poursuit donc 3 objectifs :

- Elaborer et mettre en œuvre une feuille de route sémantique (en coordination avec la gouvernance de l'interopérabilité) ;
- Créer un espace de confiance sémantique pour les acteurs de l'interopérabilité (Espace de publication des référentiels, avec une garantie de qualité, et de traçabilité) ;
- Faciliter le choix et l'utilisation des ressources sémantiques par une offre de service (alignement, moteur de recherche, traductions, algorithme d'extraction de la connaissance).

Les travaux du CGTS bénéficieront à l'ensemble de la communauté du numérique en santé :

1. Les industriels qui trouveront des ressources interopérables gratuites, facilement utilisables et accessibles ;
2. Les maîtrises d'ouvrage nationales et régionales qui auront un point d'accès central au patrimoine sémantique pour les échanges de santé ;
3. Les professionnels de santé et chercheurs qui auront des référentiels de modélisation de la connaissance, utilisables pour exploiter des données de santé (épidémiologie observationnelle ou prospective), aider à la décision, construire des ontologies de domaine au service de l'intelligence artificielle, fouiller des documents de santé (indexation/annotation/recherche).

1.2 L'écosystème des Terminologies en France

Comme vu ci-dessus, le virage numérique dans le secteur santé-social a démultiplié le besoin d'échanges de données issues de nombreux domaines de connaissance :

- **disciplines médicales** spécifiques (soins primaires, cardiologie, oncologie, maladies rares, santé mentale, etc.) ;
- **disciplines transverses** (anatomie, actes, médicaux, biologie médicale, pharmacologie ou « médicaments », etc.) ;

Ces domaines de connaissances sont représentés au sein de nombreuses ressources sémantiques qui permettent de représenter les connaissances médicales.

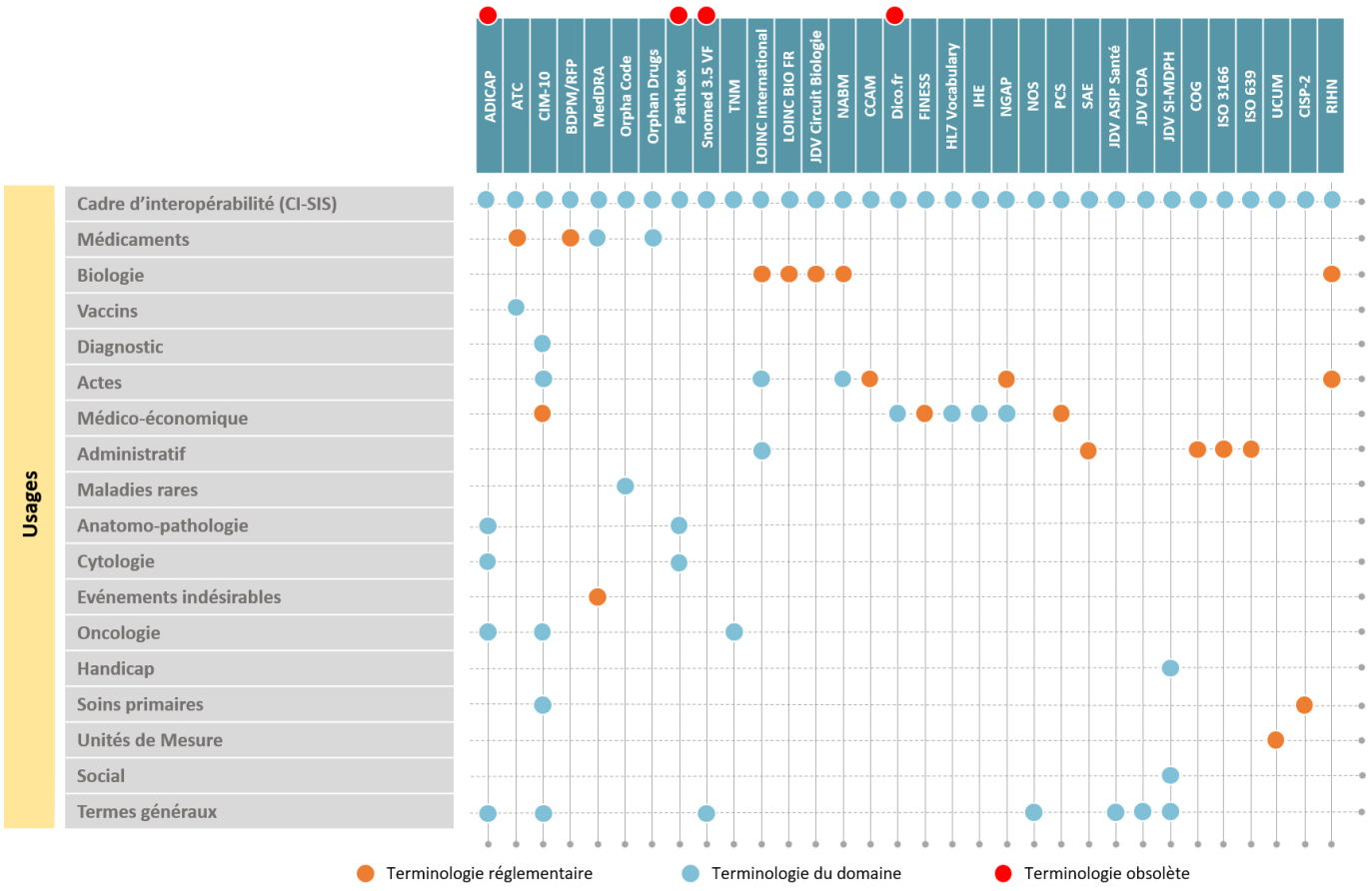
Elles sont utilisées en interface homme machine en production de soins (description des états de santé, délivrance de soins), à la coordination entre les professionnels de santé, à la facturation d'actes, etc...

Elles permettent de dématérialiser les échanges de documents de santé, d'en permettre l'interopérabilité. Les ressources sémantiques constituent un langage commun et partagé par tous les professionnels du secteur santé-social. Elles représentent ainsi le **socle nécessaire à l'interopérabilité sémantique**.

Elles permettent aussi l'exploitation des données (annotation, recherche, indexation, et extraction de la connaissance par raisonnement).

En France, le CGTS regroupe à date au sein de son Catalogue 32 Terminologies couvrant un grand nombre de domaines médicaux. Un certain nombre sont en cours d'intégration.

Figure 1 : Terminologies intégrées au Catalogue du CGTS



Les usages des ressources sémantiques sont ancrés dans la pratique des professionnels et dans le paysage des SI de santé en fonction de leur caractère obligatoire ou de leurs usages spécifiques (statistiques nationales, relevés d'indicateurs, constitution d'appels d'offres).

Ainsi, certaines Terminologies font l'objet d'un cadre réglementaire national faisant d'elles des ressources « obligatoires » pour un cas d'usage donné (CIM, CCAM, Cladimed, LOINC, MedDRA, etc.). Les Terminologies de l'OMS font partie des Terminologies que la France s'engage à utiliser.

Ainsi, la CIM-10 sert à l'analyse des causes de décès, à la production des résumés de séjours hospitaliers et à déclarer des affections de longue durée. Elle sera remplacée par la CIM-11 à court terme.

Beaucoup d'autres Terminologies sont en usage en France, auprès de professionnels de santé (ressources sémantiques de la CNAM, classification internationale des soins primaires (CISP), classification anatomique thérapeutique et chimique (ATC), nomenclature Orphanet des maladies rares, etc.).

Au total le corpus sémantique français est richement doté avec beaucoup de ressources incontournables dont il faut développer les usages.

Cependant l'offre sémantique à disposition des industriels est hétérogène en termes de qualité, d'accessibilité et d'exploitabilité.

La mise en qualité des ressources sémantiques représente un défi pour mettre en œuvre une interopérabilité sémantique.

C'est dans ce contexte (hétérogénéité de qualité et nombreuses Terminologies existantes) que SNOMED International positionne la SNOMED CT, comme terminologie multi-domaines capable de remplacer un certain nombre de Terminologies existantes.

La question de la mise à disposition d'une terminologie unique capable de rationaliser l'offre des terminologies est ainsi largement débattue en France.

1.3 La DNS a missionné l'ANS pour une étude d'évaluation de Terminologies

Dans la continuité des études menées lors des phases précédentes, **le CGTS a été missionné par la DNS pour étudier le positionnement de la SNOMED CT en France et son éventuelle adoption nationale.**

Cette étude a donc pour objectif principal d'éclairer la décision de la DNS sur le besoin d'une adoption de la SNOMED CT.

Le principe de l'étude est donc une évaluation de la SNOMED CT et de ses alternatives dans différentes situations.

2 ORGANISATION DE L'ETUDE

2.1 Les différents chantiers

Pour répondre à l'objectif de l'étude quatre axes de travail ont été définis et déclinés en chantiers :

- **Benchmark international : retour d'expérience de quatre pays**

L'objectif de cette étude est de mettre à jour les retours d'expérience (par rapport à l'étude précédente) de pays étrangers sur la mise en œuvre et les usages de la SNOMED CT. Des recommandations pour le déploiement potentiel de la SNOMED CT en France sont attendues de ces retours d'expérience.

Des entretiens ont été menés en interrogeant les **structures dont les mandats, missions et intérêts sont comparables à ceux du CGTS**. Les informations recueillies sont ainsi transposables à la gouvernance française sans biais d'intérêts privés.

Quatre pays européens ont été sélectionnés pour étudier ce retour d'expérience (Norvège ; Pays-Bas, Allemagne, Pologne). La SNOMED CT y est soit en test soit déployée.

- **Positionnement juridique de la SNOMED CT au sein d'un ensemble de ressources sémantiques d'intérêt**

Ce chantier présente un double objectif :

- Clarifier les conditions d'utilisation de la SNOMED CT associées à la licence « affiliée » qui sera en usage en cas d'adoption nationale de la SNOMED CT ;
- Positionner un panel de ressources sémantiques, dont la SNOMED CT, les unes par rapport aux autres, en fonction des droits de propriété intellectuelle accordés et de la sécurité juridique apportée par chaque licence.

Dans ce but, une analyse approfondie des droits de propriété intellectuelle accordée par chaque licence, a été réalisée. Toutes les Terminologies ainsi évaluées ont été positionnées sur une grille mesurant « ouverture des données » et « sécurité juridique ».

Afin de mener ce chantier, l'ANS a sollicité le cabinet d'avocats KGA[2].

- **Comment et pourquoi évaluer une ressource sémantique ; statut de l'évaluation de la SNOMED CT**

Compte tenu des impacts organisationnels, techniques et financiers induits par des choix de référentiels, il est important d'évaluer avant d'adoption. La question des choix méthodologiques se pose. Pour répondre à cette question, une étude bibliographique a été menée.

L'objectif de cette étude est double :

- présenter la diversité des ressources sémantiques, les divers cas d'usage , ainsi que le choix des critères et méthodes d'évaluation en fonction des différents usages ;
- positionner la SNOMED CT en fonction des données d'évaluation disponibles.

Pour mener ce chantier, l'ANS a sollicité les trois acteurs suivants :

- la société IQVIA[3] spécialiste du traitement des données de santé ;
- le LIMICS (cf. chantier scientifique ci-après) ;
- l'ISPED (cf. chantier scientifique ci-après).

[2] https://kga-avocats.fr/web/notre-identite/
[3] https://www.iqvia.com/about-us

- **Chantier Scientifique : Analyse du positionnement ou des performances de la SNOMED CT dans différents cas d'usage**

10 études représentant 10 cas d'usage ont été menées.

L'objectif est d'évaluer soit le positionnement soit les performances de la SNOMED CT pour différents cas d'usage : comparée à d'autres terminologies, la SNOMED CT répond-elle aux besoins métiers d'interopérabilité et de vocabulaire ? Assure-t-elle le rôle d'indexation, d'annotation ?

Afin de mener ce chantier, l'ANS a choisi de s'appuyer sur des équipes opérationnelles d'experts.

Le choix de 9 cas d'usage a été laissé à des experts externes à l'ANS pour bâtir l'étude sur des besoins terrain existants.

La 10ème étude a porté sur le besoin de codage et de structuration des données médicales en soins primaires. Elle a été menée par le pôle des affaires médicales et labellisation (PML) de l'ANS. Le choix de ce cas d'usage général et intéressant un très grand nombre de professionnels de santé vient en complément des autres cas d'usage spécifiques choisis par les experts. Il s'intéresse aux utilisations des Terminologies en tant qu'interface homme-machine.

Les cas d'usage ne sont pas exhaustifs, mais représentent un échantillon de situations dans lesquelles la SNOMED CT peut être utilisée :

- Interface homme machine ;
- Interopérabilité ;
- Annotation, indexation, recherche de termes ;
- Raisonnement.

L'ANS a sollicité les experts suivants :

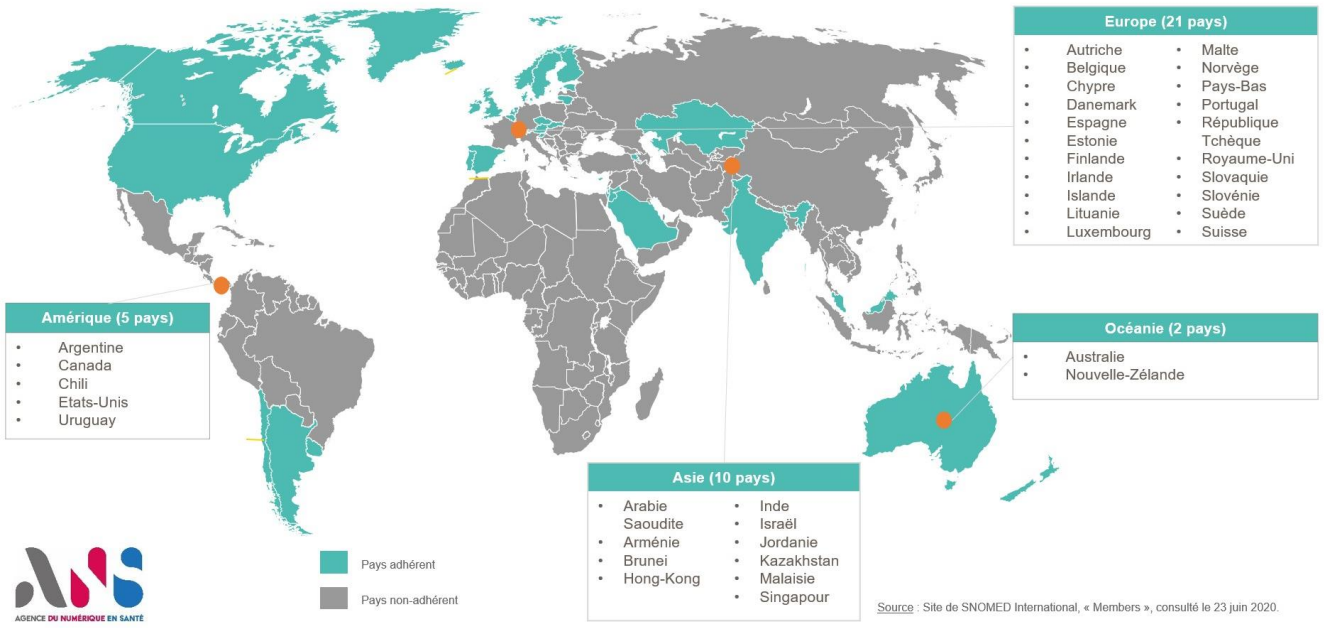
- L'ISPED : L'Institut de Santé Publique, d'Epidémiologie et de Développement est une composante de l'Université de Bordeaux rattaché au collège sciences de la santé. Spécialisé en informatique médicale, il a pour objectifs la formation, le développement d'échanges internationaux, la recherche ainsi que la valorisation de ces différentes activités. Dans le cadre de l'étude, l'ISPED évalue la performance de terminologies de référence sur les cas d'usage **oncologie** et **médicament** ;
- Le LIMICS : Le Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé est une unité de recherche interdisciplinaire. Il a pour objectifs le développement de systèmes décisionnels innovants et d'applicatifs de traitement de l'information de santé. Dans le cadre de l'étude, le LIMICS évalue la performance de différentes terminologies sur les cas d'usage **anatomie**, **génomique**, **maladies rares**, **médicament** et **recouvrement SNOMED CT – CIM-11** ;
- Le groupe DaSciM (LIX/FX Conseil) : L'équipe de Data Science and Mining (DaSciM) fait partie du Laboratoire d'Informatique de l'École Polytechnique (LIX). Elle mène des recherches en bases de données et en *Data Mining*. Dans le cadre de l'étude, le groupe s'est attaché à positionner les ressources sémantiques dans des cas d'usage d'extraction de la connaissance à partir de données textuelles aux côtés de techniques apparues récemment tel que le « *word embeddings* » (prolongement lexical) ;
- L'AP-HP (Assistance Publique - Hôpitaux de Paris) : Dans le cadre de l'étude, l'AP-HP compare la performance de plusieurs terminologies sur le cas d'usage **microbiologie.**

2.2 Ressources sémantiques comparées

2.2.1 SNOMED CT

La "Systematized Nomenclature of Medicine – Clinical Terms" (SNOMED CT ou SCT) est une terminologie descriptive internationale (38 pays membres) **multi-domaines** gérée par SNOMED International, nom commercial donné depuis 2017 à l'International Health Terminology Standards Development Organisation (IHTSDO), organisation internationale à but non lucratif fondée en 2007, basée à Londres[4] et ayant acquis les droits de la SNOMED CT au College of American Pathologist (CAP).

Figure 2 : Les 38 pays membres de SNOMED International



Elle couvre un large spectre de domaines de connaissance médicaux (elle est structurée en 19 chapitres). Elle contient environ 350 000 concepts entre lesquels existent deux millions de relations : cette ressource sémantique est une terminologie **multiaxiale** (contenant de nombreuses relations et de synonymes). Elle permet de décrire des situations variées. Les 19 chapitres de la SNOMED CT sont présentés en annexe de ce rapport mis en miroir avec les équivalences dans d'autres Terminologies.

C'est une **terminologie en constante évolution** maintenue sous la supervision de SNOMED International par un large réseau d'experts et de pays membres qui produisent en parallèle des extensions nationales ce qui lui permet de couvrir de nouveaux domaines médicaux ou d'adapter les concepts aux contextes locaux. Elle est utilisée systématiquement dans les spécifications internationales.

La SNOMED CT est une potentielle candidate au rôle de terminologie pivot[5], grâce à sa couverture multi-domaines et à sa capacité à établir des alignements sémantiques avec des terminologies de références dans leur domaine spécifique assurant ainsi une couche d'interopérabilité.

[4] https://ihtsdo.freshdesk.com/support/solutions/articles/4000095393-why-did-you-change-your-name-from-ihtsdo-to-snomed-international-# :~:text=The%20new%20company%20will%20trade,International%20is%20its%20trading%20name.
[5] Une terminologie pivot est une terminologie multi-domaines ayant la réputation d'être exhaustive, vers laquelle s'oriente la recherche de concepts en première intention.

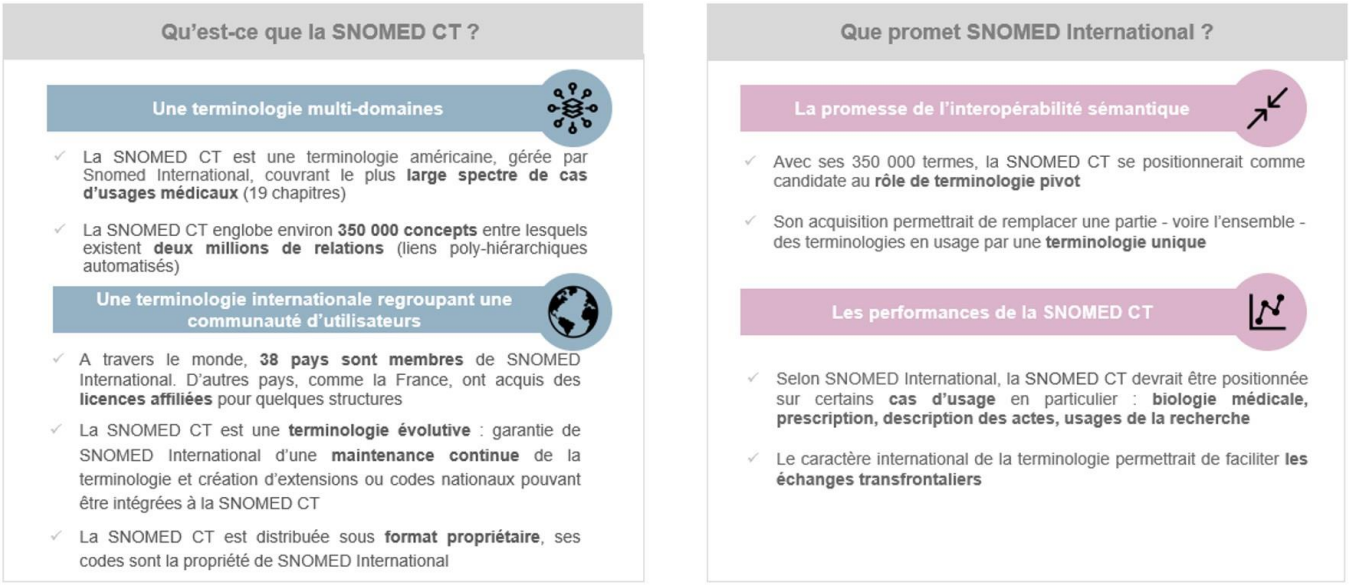
Les travaux précédents menés par l'ANS ont mis en lumière les contraintes associées à l'adoption de la SNOMED CT :

- **La SNOMED CT représente des coûts financiers importants** : adhésion à SNOMED International, prix de la licence annuelle, coûts de traduction, création d'un centre national de gestion (National Release Center – NRC) de la SNOMED CT, coûts de formation et, surtout, coûts de déploiements et d'accompagnement des professionnels. L'adhésion à SNOMED International implique les coûts suivants :
  - Droit d'adhésion calculé selon le PIB du pays (574 683 US$/507 000 €) pour la première année en France ;
  - Cotisations annuelles (du même montant de l'adhésion) ;
  - A ces coûts s'ajoutent des coûts de mise en œuvre et de maintien du NRC, des coûts de traduction, des coûts d'alignement avec les ressources sémantiques existantes en usage auprès des professionnels de santé.

- **Le déploiement de la SNOMED CT présente des enjeux organisationnels importants** : nécessité de mise en place d'un National Release Center (NRC) afin de participer à la gouvernance de SNOMED International, efforts importants pour l'intégration et le déploiement dans les organisations et cas d'usage concernés (en ville, à l'hôpital, etc.) ;
- **Le tissu industriel français ne présente pas une maturité suffisante pour l'intégration complète** de la SNOMED CT. En effet, la SNOMED CT présente un mode de fonctionnement différent des autres Terminologies. La complexité des relations entre les concepts nécessite la **production et l'intégration de jeux de valeurs adaptés**. Ainsi, la SNOMED CT constituerait un investissement lourd et de long terme pour les éditeurs dans un environnement contenant déjà de nombreuses terminologies qu'il conviendrait de maintenir en parallèle.

La figure ci-après résume la présentation de la SNOMED CT :

Figure 3 : Présentation de la SNOMED CT



Répondre à la question d'une adoption nationale de la SNOMED CT revient donc à déterminer si la SNOMED CT représente une opportunité pour structurer l'écosystème des Terminologies en France.

La SNOMED CT a été étudiée au cours de cette étude après souscription de licences affiliées auprès de SNOMED International. La traduction de la SNOMED a été nécessaire pour étudier cette terminologie sur des données françaises des entrepôts de données des CHU de Bordeaux et de Rouen (cf. infra).

2.2.2 Ressources sémantiques de comparaison

Un panel de 28 ressources sémantiques a été étudié dans les différents chantiers de l'étude pour être comparé ou positionné relativement à la SNOMED CT.

19 ressources ont été étudiées pour le chantier juridique. Il s'agit de Terminologies du Catalogue (Orpha, LOINC, Medicabase, BDPM, MedDRA, Cladimed) dont les licences devaient être clarifiées ou de terminologies en cours d'évaluation pour des cas d'usages spécifiques : NCBI Taxonomy pour les microorganismes, FMA, Uberon pour l'anatomie, HGNC, Gene Ontology, HPO pour la génomique, ChemidPlus et NUVA pour les médicaments substances chimiques toxiques et vaccins.

22 Terminologies ont été utilisées dans les études du chantier scientifique.

Toutes ces ressources sont détaillées dans les différents rapports annexes liés au chantier scientifique.

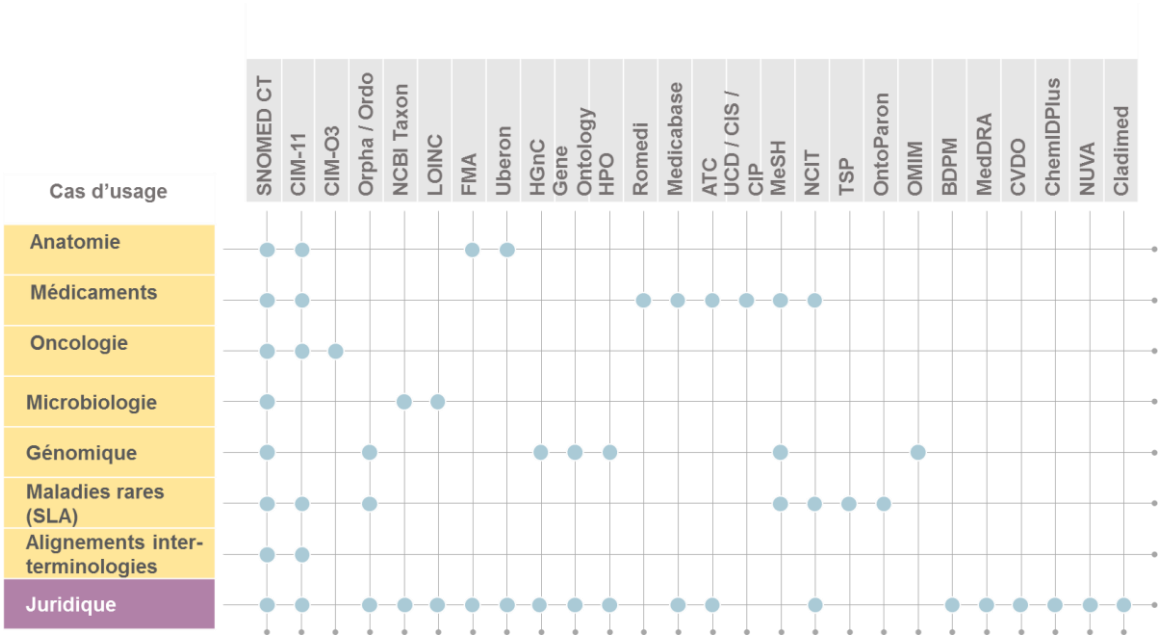
La CIM-11, CIM-O3 et l'ATC font partie des classifications de l'OMS[6]. Elles couvrent les domaines médicaux, oncologie et médicament.

La CIM-11 a été librement téléchargée à partir des API de l'OMS. Les études ont été menées à partir de la fondation ontologique de la CIM-11[7] qui contient environ 130 000 termes associés à une URI dont 60000 termes préférés. Comme pour la SNOMED CT, une traduction a été nécessaire (cf. Infra). La composante ontologique ("fondation", "*foundation*" en anglais) contient tous les concepts, les relations hiérarchiques de la CIM-11 et la grammaire de post-coordination.

A noter que la CIM-11 complète (fondation) ne doit pas être confondue avec la linéarisation morbi mortalité de la CIM-11 qui contient environ 33 000 entités codées au niveau des catégories avec un identifiant OMS[8] Il s'agit d'un sous-ensemble de la composante ontologique similaire à la structure de la CIM-10 : hiérarchie mono-axiale, sans les entités détaillées et les vues spéciales ("*special views*").

La figure ci-après synthétise les différentes ressources sémantiques de comparaison et les cas d'usage associés.

*F*igure 4 : Terminologies analysées dans le cadre de l'étude



[6] https://www.who.int/classifications/en/
[7] https://icd.who.int/icdapi
[8] https://icd.who.int/browse11/l-m/en
La hiérarchie de la CIM-11 est organisée en chapitre/blocs/sous blocs/catégories et sous catégories.
La fondation contient l'ensemble des termes de la CIM-11 ainsi que leurs liens hiérarchiques et sémantiques. Les linéarisations sont des extraits de la fondations définis et organisés dans un but précis (description des morbi mortalités, des atteintes cancéreuses, etc.). La fondation contient environ 60 000 concepts et 130 000 termes (mars 2020).

2.2.3 La problématique de la traduction

Cette étude a posé la problématique de la traduction des ressources sémantiques en français.

En effet, pour évaluer une ressource sémantique dans des cas d'usage français et a fortiori pour l'adopter, il est indispensable d'en disposer une version française.

Cette problématique s'est posée de manière aiguë pour la SNOMED CT et la CIM-11 pour lesquelles les versions disponibles étaient en anglais ou espagnol.

Pour pouvoir analyser les contenus de ces Terminologies en action, il a été nécessaire de les traduire en français.

L'ANS s'est servie d'un l'outil de traduction automatique par « neural machine Translation » (NMT) développé par FX Conseil (Ecole Polytechnique, Saclay) pour produire des versions françaises (base line) de qualité suffisante pour mener l''étude.

Dans cette tâche l'ANS a également été activement supportée par l'équipe du Département d'Informatique et d'Information Médicales (D2IM) LIMICS du CHU de Rouen.

L'outil de traduction, le « workflow » et les corpus qualifiés multilingues sont dans une boucle d'amélioration continue pour maintenir et améliorer les traductions.

Ce travail de traduction a fait l'objet de publications au congrès de L'OMS[9] (WHO FIC network 2020) et au congrès du LOUHI 2020[10] (Health Text Mining and Information Analysis).

[9] https://www.who.int/classifications/network/whoficnetworkannualmeeting/en/
[10] https://lll.cnrs.fr/evenement/louhi-2020-the-11th-international-workshop-on-health-text-mining-and-information-analysis/

3 RESULTATS

3.1 Retours d'expérience (Benchmark) international

3.1.1 Les pays interrogés

L'étude s'est appuyée sur le retour d'expérience de 4 pays représentatifs de 4 cas de figure[11] :

- **Pays-Bas** : pays fondateur et adhérent de SNOMED International ;
- **Allemagne** : pays non-adhérent de SNOMED International et mettant en œuvre un test de la SNOMED CT sur des échanges de données inter-hospitalières ;
- **Norvège** : pays adhérent à SNOMED International également dans une phase expérimentale de la SNOMED CT ;
- **Pologne** : pays adhérent depuis 2011 et ayant résilié son adhésion à SNOMED International en 2019.

Des entretiens téléphoniques ont été menés avec des représentants des homologues internationaux du CGTS, ayant les mêmes obligations, périmètres de missions et prérogatives que le CGTS.

Les détails de l'étude sont accessibles dans **l'annexe P1.0.**

3.1.2 Les enseignements

3.1.2.1 Les cas d'usages couverts par SNOMED CT dans les pays interrogés

Les cas d'usage dans lesquels la SNOMED CT a été déployée sont les suivants :

1. La SNOMED CT ne remplace pas les terminologies existantes, elle est **toujours maintenue dans un écosystème de terminologies** (NL, NO, DE, CH). Un cas d'usage n'est jamais entièrement codé avec la SNOMED CT, qui doit être complétée par d'autres terminologies**.**
2. **Les principaux domaines** où elle a fait ses preuves sont : *microbiologie, anatomie, diagnostic, médicaments, recherche et innovation (entrepôts de données) (NL, NO, DE).* Les pays membres (NL, NO) identifient **peu de cas d'usage où la SNOMED CT n'a pas pu être déployée.** C'est le cas pour les soins infirmiers et génomique (NL, CH). Certains cas d'usage, comme le médicament, ont été développés avec succès dans certains pays (NO, CH), mais ont rencontré plus de difficultés dans d'autres (NL).
3. **Des extensions nationales de la SNOMED CT sont nécessaires** (NL, CH), et sont maintenues par le centre de compétence SNOMED CT (NRC) local si non-intégrées à la SNOMED CT. En effet, la SNOMED CT ne peut couvrir tous les domaines de connaissance et surtout les spécificités locales. Il est fréquent que des pans entiers (Chapitres) de la SNOMED CT (par exemple médicaments aux Pays-Bas) ne soient pas utilisés.

[11]Sur demande spécifique, la Suisse (e-health Suisse) et les Etats-Unis (NLM et CDC), tous deux adhérents de SNOMED International, ont été contactés dans une seconde phase d'entretiens. e-health Suisse a pu répondre aux sollicitations de l'ANS, le compte-rendu des échanges figurant en annexe de ce rapport. Le CDC a pu fournir ses récentes doctrines d'adoption de Terminologies. De son côté, en date de rédaction du rapport, la NLM a accusé réception des demandes d'échanges, se montrant disposée à un entretien r au moment où le monde entier s'est confiné pour cause de COVID-19.

3.1.2.2 Les éléments à prendre en compte lors du déploiement de la SNOMED CT

Les éléments à prendre en compte lors du déploiement de la SNOMED CT sont les suivants :

1. **L'implication des industriels est l'un des facteurs clefs de succès dans le déploiement de la SNOMED CT**, et nécessite qu'une feuille de route soit clairement définie dans la stratégie publique (NO, PL). Ainsi, il est préférable **de communiquer très rapidement sur une adhésion complète à SNOMED International** afin d'offrir cette stabilité (NO) ;
2. La SNOMED CT ne peut être déployée **sans stratégie nationale sur le numérique en santé (NL, PL)** ;
3. Le succès du déploiement d'une terminologie telle que la SNOMED CT repose sur une mise en œuvre progressive (soit par cas d'usage, soit géographiquement) associée à un accompagnement des acteurs. Il est donc important de sélectionner les cas d'usages les plus adaptés, de déployer par expérimentation avant d'en généraliser l'emploi (NO, GE, CH) ;
4. La SNOMED CT représente un **investissement financier à long terme, parfois conséquent (PL) et requiert un effort d'intégration important au sein des logiciels métiers (NO)** ;
5. La SNOMED CT requiert un **effort de déploiement important** du fait de son organisation particulière (centre de compétence (NRC), traduction, intégration, production de jeux de valeurs) au dépend de l'intégration d'autres terminologies comme la CIM-11 par exemple (PL,) ;
6. **Un effort de traduction est nécessaire (NL, GE, CH).** Cette traduction progresse au fur et à mesure du développement de cas d'usage. Les traductions ne sont pas disponibles avant l'achat de la SNOMED CT. Des phases d'expérimentation permettent de vérifier la présence des concepts traduits dans les dossiers patients nationaux.

Au total, ces enseignements réaffirment les conclusions 1, 3, et 4 de l'étude ASSESS CT 2017 menée par Europe sur l'évaluation de la SNOMED CT et de ses conditions d'adoption[12] :

- Toute décision concernant l'adoption et le rôle des ressources terminologiques, y compris SNOMED CT, doit faire partie d'une stratégie plus large, cohérente et axée sur les priorités pour optimiser les bénéfices de l'interopérabilité sémantique dans les données de santé, et de la stratégie globale de e-santé de l'Union européenne et ses États membres ;
- SNOMED CT devrait faire partie d'un écosystème de terminologies, y compris les terminologies internationales (par exemple, la famille des classifications de l'OMS) et les terminologies d'interface utilisateur, qui prennent en compte le multilinguisme en Europe et la communication clinique à travers un langage professionnel multidisciplinaire et un langage profane ;
- L'adoption de SNOMED CT devrait être réalisée progressivement plutôt que d'un seul coup, en développant des jeux de valeurs qui répondent aux exigences d'interopérabilité pour les cas d'utilisation prioritaires, et en étendant ces ensembles sur plusieurs années.

[12] http://assess-ct.eu/index.php?id=start0

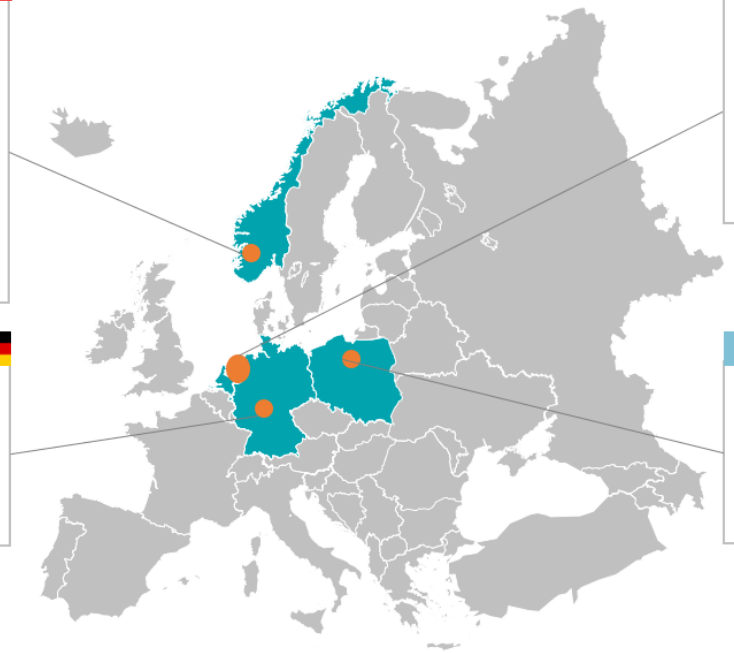
Figure 5 : Etat des lieux de l'adoption de la SNOMED CT dans les pays européens étudiés

Norvège - Directorate of eHelse

- **Phase expérimentale** sur des cas de diagnostic et dossier patient électronique (Epic) dans un grand nombre de cas d'usage.
- Résultats satisfaisants de la SCT sur **les cas d'usage** suivants : anatomie, diagnostic et médicaments. **Aucun cas d'usage n'a pu être entièrement codé avec la SCT.**
- Choix stratégique de la Norvège **d'une adhésion complète** suite à la phase expérimentale afin d'offrir une visibilité au tissu industriel.

Allemagne - Dimdi

- Adoption de plusieurs **licences affiliées**, pour tester la SCT sur des données hospitalières en exploitation (projet universitaire).
- En cas d'adhésion, **maintien de l'écosystème actuel des terminologies** (OMS, facturation, médicaments),
- Pas d'adhésion immédiate à Snomed International.



Pays-Bas – Nictiz

- Intégration de la SCT dans un **écosystème de terminologies.**
- Très fort niveau d'intégration de la SCT, bien **qu'aucun cas d'usage n'ait pu être entièrement codé** avec la SCT (cas de la microbiologie).
- Création d'**extensions nationales** de la SCT, maintenues par le NRC.

Pologne - CSIOZ

- **Absence d'intégration de la SCT dans les logiciels métiers** en huit ans d'adhésion.
- Complexité d'intégration et d'usage pour le tissu industriel.
- Contrainte financière de la SCT pour la Pologne.

3.1.3 Résumé : ce que ce chantier nous apprend sur le déploiement et les usages de la SNOMED CT

Tableau 1 : Chantier retour d'expérience international

Avantages	Inconvénients
- La SNOMED CT a fait ses preuves sur un grand nombre de domaines : microbiologie, anatomie, diagnostic, médicaments, recherche et innovation (entrepôts de données) (NL, NW, DE). - La SNOMED CT est adoptée par 38 pays et utilisée dans plus de 50 pays, assurant sa maintenance en continu par une communauté internationale.	- La SNOMED CT ne remplace pas les terminologies existantes. Elle doit être maintenue dans un écosystème de terminologies. - Des extensions nationales de la SNOMED CT sont nécessaires (NL), et sont maintenues par le NRC local si non-intégrées à la SNOMED CT. - La SNOMED CT représente un investissement financier à long terme conséquent (PL) et requiert un effort d'intégration (NW).

3.2 Evaluation juridique

3.2.1 Objectifs de l'évaluation juridique

Pour rappel, cette étude a pour double objectif :

- Clarifier les conditions d'utilisation de la SNOMED CT associées à la licence affiliée qui sera en usage en cas d'adoption nationale de la SNOMED CT ;
- Positionner un panel de ressources sémantiques, dont la SNOMED CT, les unes par rapport aux autres, en fonction des droits de propriété intellectuelle accordés et de la sécurité juridique apportée par chaque licence.

Ceci permettra d'apporter un éclairage juridique sur l'adoption éventuelle de la SNOMED CT et son positionnement par rapport à la doctrine d'ouverture des données publiques à laquelle l'ANS souhaite se conformer.

Ces travaux juridiques ont été réalisés par le cabinet d'avocats KGA[13] et le service juridique de l'ANS. Le détail de l'étude est exposé en **annexe P2.0**.

Les terminologies pointées par le Catalogue des Terminologies de santé du CGTS sont issues d'unités de production rattachées à des personnes morales variées. Elles sont des **œuvres numériques** protégées par des droits de propriété intellectuelle.

3.2.2 Contexte juridique de la publication de ressources sémantiques par le CGTS

« La France porte une tradition de transparence démocratique et de partage des informations détenues par la puissance publique. Dans le droit fil de cette tradition, une politique ambitieuse a été engagée depuis 2015, notamment en matière d'ouverture des données publiques »[14].

Dans cette logique la France impose la diffusion des données publiques en open data (ou ouverture des données).

L' « open data » peut être défini comme la mise à disposition en ligne de données publiques[15], librement réutilisables par toute personne. Cela nécessite notamment que les données soient mises à disposition dans un format ouvert, facilement exploitable par une machine.

L'ouverture des données publiques contribue notamment à :

- améliorer le fonctionnement démocratique, par la transparence, la concertation et l'ouverture à de nouveaux points de vue ;
- améliorer l'efficacité de l'action publique ;
- proposer de nouvelles ressources pour l'innovation économique et sociale : les données partagées trouvent des utilisateurs qui les intègrent dans de nouveaux services à forte valeur ajoutée économique ou sociale[16].

Le régime juridique de publication de ressources sémantiques doit être clair pour ne pas créer d'insécurité pour les utilisateurs.

Dans ce cadre, l'open data peut être vu comme un outil pour apporter sécurité juridique et transparence aux utilisateurs de ressources sémantiques.

En France, la loi pour une république numérique (LRN)[17] du 7 octobre 2016 a instauré l' « open data » par défaut en imposant la diffusion en ligne de différentes catégories de données publiques. L'objectif affiché de cette loi était de préparer « le pays aux enjeux de la transition numérique et de l'économie de demain. Elle promeut l'innovation et le

[13] Cabinet d'avocat d'affaire (https://kga-avocats.fr/web/) 44, avenue des Champs-Elysées - 75008 Paris.
[14] https://www.gouvernement.fr/action/l-ouverture-des-donnees-publiques
[15] Données produites ou reçues dans le cadre de l'exécution d'une mission de service public par une entité publique ou privée en charge d'une telle mission.
[16] « L'ouverture des données » - 15 mai 2017 – gouvernement.fr (https://www.gouvernement.fr/action/l-ouverture-des-donnees-publiques)
[17] Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique

développement de l'économie numérique, une société numérique ouverte, fiable et protectrice des droits des citoyens. Elle vise également à garantir l'accès de tous, dans tous les territoires, aux opportunités liées au numérique »[18].

L'ANS, en tant que groupement d'intérêt public chargé d'une mission de service public, est soumise aux dispositions relatives à l'ouverture des données publiques de la loi pour une république numérique (LRN).

Ainsi, l'ANS est notamment tenue de publier :

- les bases de données qu'elle produit ou reçoit et qui ne font pas l'objet d'une diffusion publique par ailleurs ;
- les données dont la publication présente « un intérêt économique, social, sanitaire ou environnemental », etc.[19].

Les ressources sémantiques diffusées par l'ANS correspondent à un ensemble de classifications, de nomenclatures, de jeux de valeurs, de terminologies, d'ontologies, d'alignements sémantiques etc.[20] Ces ressources sont la propriété d'unités de production aux statuts variés susceptibles d'impacter la qualification de leur contenu. L'ensemble de ces ressources sémantiques contenues au sein du catalogue du Centre de Gestion de Terminologies de Santé (CGTS) constitue à la fois :

- Une base de données ;
- Un document administratif[21].

3.2.3 Méthodologie

Pour mener l'évaluation juridique, une grille d'analyse a été construite en collaboration avec le cabinet KGA Avocats.

Les licences de ressources sémantiques sont analysées selon deux axes :

- la sécurité juridique pour les utilisateurs ;
- l'ouverture des données (la conformité à l'open data).

Un panel de 19 ressources sémantiques dont la SNOMED CT a été sélectionné pour l'étude.

Ces ressources ont été choisies parmi celles du catalogue du CGTS dont les droits de propriété intellectuelle devaient être clarifiés (MedDRA, CIM-11, ATC, Medicabase, Cladimed, LOINC) ou dont l'adoption est en cours d'évaluation (SNOMED CT, HPO, HGNC, NUVA, NCBI Taxonomy, Uberon, NCIT, BDPM, Gene Ontology).

Dans le cas de la SNOMED CT, l'analyse des différentes licences (nationale, affiliées, recherche) a été menée par le cabinet d'avocats KGA, à l'aide d'entretiens avec SNOMED International.

Toutes les licences ont été évaluées au travers de la grille d'analyse.

3.2.4 Entretien avec SNOMED International

L'entretien avec SNOMED International a permis de préciser les conditions d'adhésion/résiliation à SNOMED International ainsi que les différents types de licences concédées par l'organisation SNOMED International.

3.2.4.1 Adhésion à SNOMED International

L'adhésion à SNOMED International est payante (frais d'adhésion puis abonnement) et implique de désigner au sein d'un Etat membre (ou membre) :

- un organisme désigné « membre » représentant de l'Etat à la gouvernance de SNOMED International ;
- un National Release Center (NRC), en charge de la diffusion de la SNOMED CT dans le pays membre.

[18] https://www.economie.gouv.fr/republique-numerique - 10 octobre 2016
[19] Article L 312-1-1 du code des relations entre le public et l'administration
[20] « Espace des terminologies de santé / Ressources », site Internet de l'ASIP Santé, consulté le 28 octobre 2019.
[21] Article L 300-2 du code des relations entre le public et l'administration

Le pays membre obtient ainsi le droit :

- **d'utiliser la version internationale de la SNOMED CT**, de l'incorporer dans ses produits et services et de distribuer ces derniers. Les licences Affiliées concédées aux utilisateurs sont directement signées avec SNOMED International ;
- **de créer des versions modifiées de la SNOMED CT**, en particulier, des versions nationales (traduction de la SNOMED CT), ainsi que des extensions et le droit d'utiliser de modifier ceux-ci ;
- **de publier sa version nationale de la SNOMED CT** et de concéder des licences spécifiques sur cette version nationale (dont les termes ne peuvent pas restreindre les obligations imposées par la licence Affiliée), aux utilisateurs de son territoire ainsi qu'aux membres de SNOMED International.

3.2.4.2 Résiliation de l'adhésion à SNOMED International

La résiliation de l'adhésion à SNOMED International par un membre a des conséquences sur l'utilisation de la SNOMED CT et les droits des extensions nationales éventuellement créées par le membre.

- ***Utilisation de la SNOMED CT***

Dans l'hypothèse où un membre déciderait de résilier son contrat avec SNOMED International, ce dernier contacte l'ensemble des organismes ayant souscrit une Licence Affiliée, afin de leur proposer de leur facturer directement les frais de licence liés à l'utilisation de SNOMED CT.

Les Affiliés devront prendre directement à leur charge les frais d'utilisation de la SNOMED CT. Dans le cas contraire, leurs droits et ceux des sous-licenciés sont révoqués. Ils sont tenus de retirer de leurs systèmes d'information toutes les copies de la SNOMED CT et, sur demande, pouvoir certifier par écrit que cette suppression a bien été réalisée.

- ***Droit de propriété intellectuelle des extensions nationales***

Après résiliation définitive du contrat d'un membre, les droits de propriété intellectuelle des extensions nationales qu'il a créées sont transférés à un organisme qu'il a désigné. Les statuts de SNOMED International ne précisent pas si l'organisme désigné doit adhérer à son tour à SNOMED International.

Si le gouvernement ne désigne pas de nouvel organisme, les droits de propriété intellectuelle seront cédés à l'organisme choisi par SNOMED International ou à SNOMED International directement et le membre conserve uniquement un droit d'utilisation précaire[22] de la version en vigueur au moment de la résiliation.

3.2.4.3 Les licences de la SNOMED CT (version décembre 2019-février 2020)

L'organisation SNOMED International propose trois différents types de licences permettant un usage de la SNOMED CT. Chacune confère des droits différents et est fonction de l'adhésion du pays à SNOMED International.

[22] Ce droit d'utilisation est résilié dès qu'un nouvel organisme est désigné et il est limité conformément à la clause 6.2 des statuts de SNOMED International.

Figure 6 : Différentes licences mises à disposition par SNOMED International

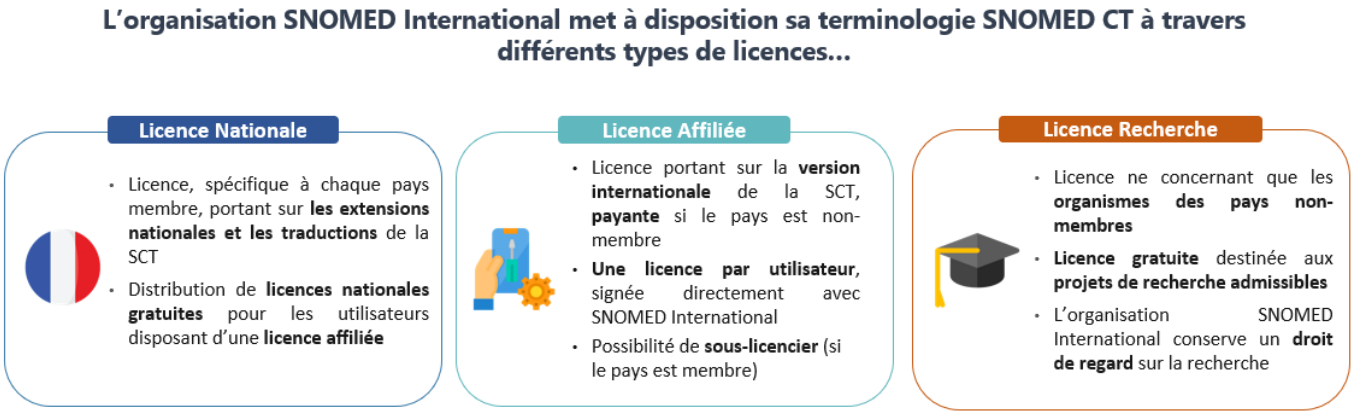
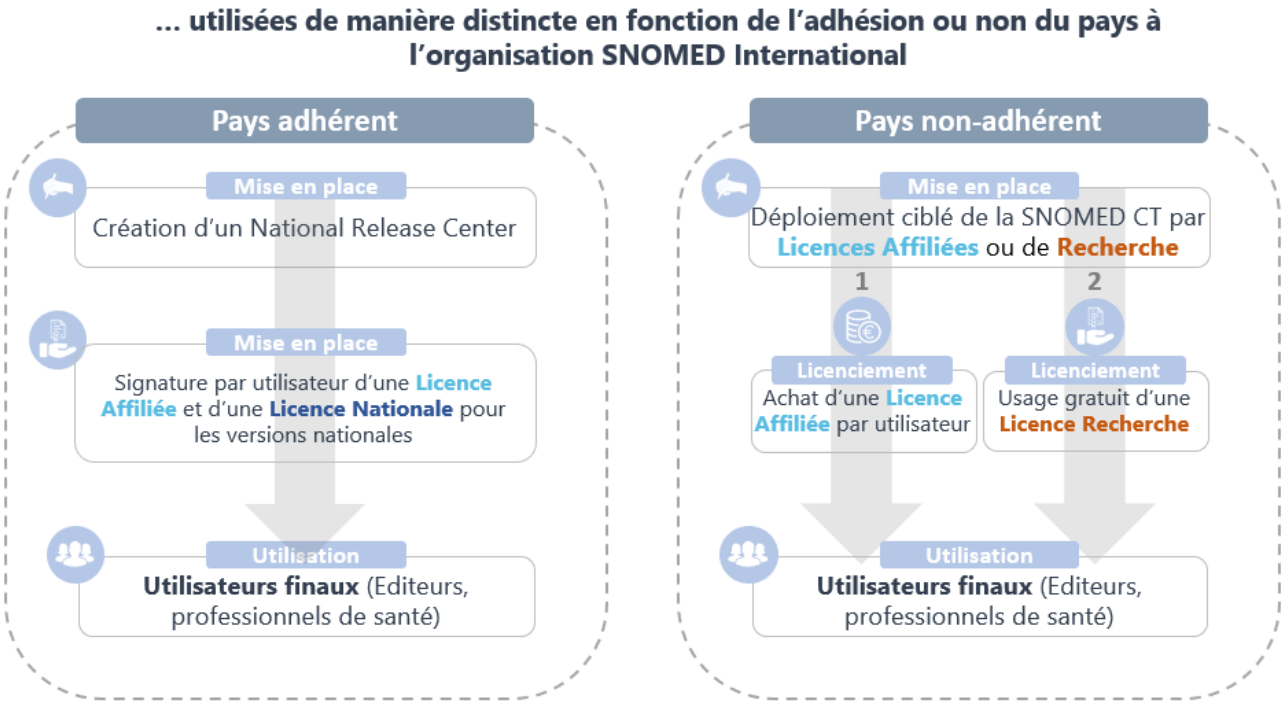


Figure 7 : Utilisation des licences SNOMED CT dans les cas de figure pays adhérent pays vs non-adhérent



Les utilisateurs d'un pays membre signent les licences affiliées avec SNOMED International par le biais du NRC. Cette licence offre aux utilisateurs le droit d'utiliser la version internationale de la SNOMED CT ainsi que concéder des sous-licences d'utilisation de la SNOMED CT (un utilisateur final, tel qu'un professionnel de santé, utilisant la SNOMED CT *via* une solution qui l'intègre n'a pas besoin d'obtenir une licence). Les utilisateurs doivent également signer une licence nationale pour l'utilisation des versions nationales.

Au sein d'un **pays non-membre**, chaque utilisateur est libre d'acquérir individuellement une licence affiliée lui offrant un usage de la ressource sémantique. L'organisation SNOMED International propose également des licences de recherche gratuites, destinées aux projets de recherche sélectionnés par SNOMED International[23] et offrant un usage de la SNOMED CT restreint.

[23] Alinéas 93 -107 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)

L'analyse de la licence affiliée de la SNOMED CT par le cabinet d'avocats KGA démontre une contradiction entre les règles imposées par SNOMED International et l'open data, ainsi qu'une certaine insécurité juridique (cf. annexe P2.0) :

- **un utilisateur ne peut pas extraire une partie substantielle de la SNOMED CT à partir des systèmes de l'affilié[24] ;**
- **pas d'autorisation de modifications de la SNOMED CT Core (impossibilité de créer des jeux de valeurs mixtes multi terminologies, impossibilité d'élimination des codes SNOMED CT non opérationnels)[25] ;**
- **insécurité en matière d'interopérabilité en cas de résiliation[26].**
- **SNOMED International ne garantit pas être titulaire des droits qu'elle concède, malgré le fait que SNOMED CT ne soit pas une source primaire et qu'elle intègre d'autres ressources sémantiques[27] ;**
- **la Licence concède à l'organisme Affilié des droits (utilisation, modification, traduction, etc.) qui ne sont pas précisément définis et qui peuvent être uniquement exercés dans des cas spécifiques (usage interne, projet de recherche, etc.)[28].**

3.2.5 Les ressources sémantiques sont principalement publiées via des licences ouvertes

L'analyse juridique des licences ci-dessus, selon les dimensions ouverture des données et sécurité juridique pour les utilisateurs, permet de les cartographier. Le graphe ci-après illustre cette analyse.

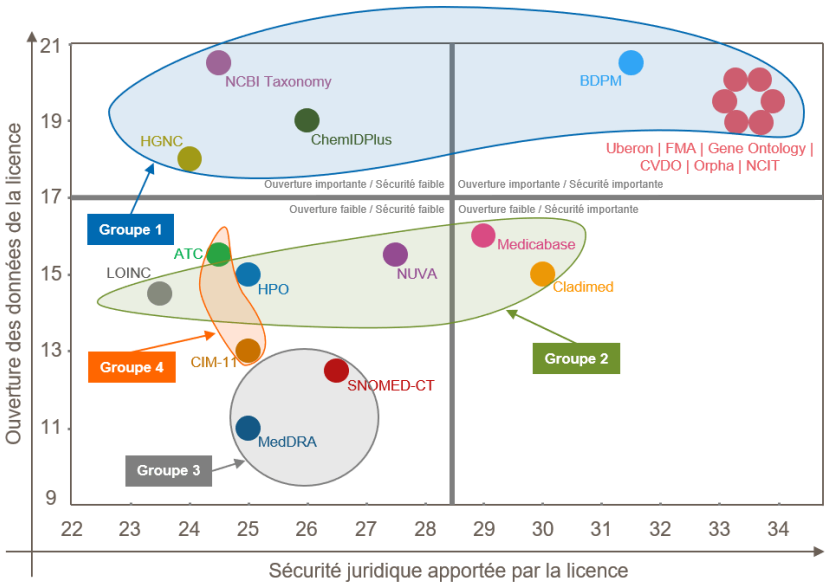
- L'**axe vertical** est relatif au paramètre « **Ouverture des données** » (libre accessibilité et exploitabilité des licences pour les utilisateurs des ressources sémantiques) ;
- L'**axe horizontal** est relatif à la « **Sécurité juridique** » des licences (la clarté des stipulations offrant un cadre juridique clair aux utilisateurs).

La moyenne des notes d'ouverture obtenue est de **17/22,5** ; celle relative à la sécurité juridique est de **28,5/34,5**.

Les axes passant par ces coordonnées déterminent 4 cadrans égaux : dans les zones supérieures se trouvent les licences ouvertes, dans les zones de droite se trouvent les licences qui apportent le plus de sécurité juridique.

Les licences optimales (ouvertes et claires) se trouvent donc dans le cadran supérieur droit.

Figure 8 : Classement des licences terminologiques en fonction de l'ouverture des données et la sécurité juridique



[24] Alinéa 57 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)
[25] Alinéas 63, 64 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)
[26] Alinéas 76 à 79 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)
[27] Alinéas 57-59 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)
[28] Alinéa 61 du mémorandum d'analyse de la SNOMED CT (KGA Avocats)

Au regard de l'ouverture des données, quatre groupes peuvent être définis.

3.2.5.1 Groupe 1 : groupe des licences ouvertes

Dix des dix-neuf ressources sémantiques étudiées, sont diffusées sous licences ouvertes, sans restriction d'utilisation pour les utilisateurs. Ces licences leur confèrent un droit non-exclusif de réutilisation des données (usage libre et gratuit de la ressource sémantique, possibilité de reproduire, modifier la ressource sémantique, de créer des œuvres dérivées et de l'utiliser à des fins commerciales). Il s'agit des ressources sémantiques suivantes : NCBI Taxonomy, ChemIDPlus, Uberon, Gene Ontology, Orpha, NCIT, CVDO, FMA, BDPM, HGNC et Orpha.

3.2.5.2 Groupe 2 : libre diffusion

Cinq ressources sémantiques sont soumises à des licences de libre diffusion qui autorisent la redistribution de la ressource sémantique, mais qui restreignent les droits concédés (pas de modification possible, pas de création d'œuvre dérivée, etc.) : HPO, NUVA, Medicabase, Cladimed et LOINC.

3.2.5.3 Groupe 3 : licences propriétaires

Deux des dix-neuf ressources sémantiques du panel (MedDRA et SNOMED CT) sont diffusées selon des modèles propriétaires et obtiennent les scores d'ouverture les plus bas.

La SNOMED CT, est diffusée sous une licence payante, qui ne permet pas de créer des œuvres dérivées librement et limite la libre exploitabilité pour les utilisateurs. Elle présente par ailleurs des conditions de résiliation très contraignantes en contradiction avec les principes de l'open data.

MedDRA est classée dernière en termes d'« Ouverture des données », car sa licence accorde un droit d'usage exclusif, pour une durée limitée (1 an) et n'autorise pas une utilisation libre de la ressource sémantique (usage interne, interdiction de modification, limitation des extractions de la ressource sémantique sous forme de fichier complet et de sous-licenciement, etc.). Ses termes limitent son utilisation à un monde d'experts de la pharmacovigilance devant utiliser cette ressource.

3.2.5.4 Groupe 4 : ressources OMS

Les ressources sémantiques de l'OMS, ATC et CIM-11, ne peuvent pas être diffusées librement et leur utilisation est restreinte (pas de modification autorisée sans l'accord de l'OMS avec le respect de la procédure de demande de modification, ou utilisation uniquement à des fins de recherche). Ces licences sont contraires à l'Open data. Les licences des ressources sémantiques de l'OMS (ATC et CIM-11) devraient, selon les déclarations de l'OMS, faire l'objet d'une mise à jour afin d'en permettre un libre usage (y compris commercial) avec pour seules limitations que les modifications doivent être approuvées par l'OMS et que les versions modifiées ont l'obligation de mentionner cette approbation[29].

3.2.5.5 Sécurité juridique

Environ la moitié des ressources sémantiques étudiées (neuf sur dix-neuf) offrent un cadre juridique clair et fiable (positionnement à droite de la figure) pour les utilisateurs, limitant les risques de mauvaises interprétations du cadre juridique de la réutilisation des données. Il s'agit de : BDPM, Uberon, FMA, CVDO, NCIT, Orpha, Gene Ontology, Medicabase et Cladimed.

L'autre moitié des ressources sémantiques (10 sur 19) est associée à des licences positionnées à gauche de la figure. Elles présentent des imprécisions, pouvant donner lieu à des interprétations juridiques erronées ou équivoques. Par exemple, elles ne permettent pas de définir précisément les droits concédés ou les conditions de leur jouissance. On trouve dans ce groupe les ressources sémantiques suivantes : NCBI Taxonomy, HGNC, ChemIDPlus, NUVA, HPO, ATC, LOINC, SNOMED CT, CIM-11, MedDRA.

[29] Courrier de l'OMS au Centre collaborateur français de l'OMS reçu le 23/04/2020.

3.2.6 Discussion et conclusion

Cette étude a démontré une tendance à l'ouverture des données dans l'écosystème des ressources sémantiques en santé.

Cette tendance est cohérente avec la logique d'open data adoptée par la France et est en accord avec la constitution d'un corpus national, librement accessible et librement exploitable destiné à servir de langage commun pour les échanges dans le secteur de la santé.

L'ouverture des données permet d'apporter la sécurité juridique et la transparence aux utilisateurs de la donnée.

Dans le cadre du choix d'un langage commun pour l'échange de données de santé, il est préférable de privilégier une ressource sémantique ouverte ou *a minima* soumise à une licence permettant une libre diffusion.

En cas d'adoption d'une ressource sémantique avec une licence propriétaire, comme la SNOMED CT, il faudrait prévoir des modalités de réversibilité dans la licence d'utilisation, afin de permettre une transition vers un autre système sémantique sans pénaliser le tissu industriel si, à terme, la licence venait à être résiliée.

3.2.7 Résumé : ce que ce chantier nous apprend sur le positionnement juridique de la SNOMED CT

Tableau 2 : Chantier juridique

Avantages	Inconvénients
- L'organisme SNOMED International propose une licence affiliée clairement définie pour sa terminologie. - Un outil MLDS permet la distribution des licences dans les pays membres. - Le pays membre a la possibilité de modifier sa version et proposer des modifications de la version internationale (extensions, traductions) de la SNOMED CT.	- La SNOMED CT est diffusée via un modèle fermé, limitant la libre exploitabilité de la terminologie pour les utilisateurs. - A ce titre, la SNOMED CT fait figure d'exception dans un écosystème où les ressources sémantiques sont diffusées majoritairement sous format ouvert. - Les conditions de résiliation à SNOMED International sont très contraignantes : les utilisateurs doivent payer individuellement une licence affiliée pour conserver les codes de la SNOMED CT. - Ambiguïté de certaines clauses de la licence affiliée.

3.3 Comment et pourquoi évaluer une ressource sémantique, statut de l'évaluation de la SNOMED CT (données bibliographiques)

Un travail bibliographique a été réalisé, dans le cadre de cette étude, centré sur l'évaluation des ressources sémantiques et sur celle de la SNOMED CT en particulier.

L'objectif de cette étude est multiple :

- présenter la diversité des ressources sémantiques, les divers cas d'usage ainsi que le choix des critères et méthodes d'évaluation en fonction des différents usages ;
- positionner la SNOMED CT en fonction des données d'évaluation disponibles.

Ce rapport vise également à familiariser le lecteur avec les différentes ressources sémantiques.

L'étude a été effectuée sur la base Medline via le moteur Pubmed. Le détail de l'étude est présenté **en annexe P3.0.**

Les ressources sémantiques sont variées allant de la simple liste de termes à une ontologie formellement définie.

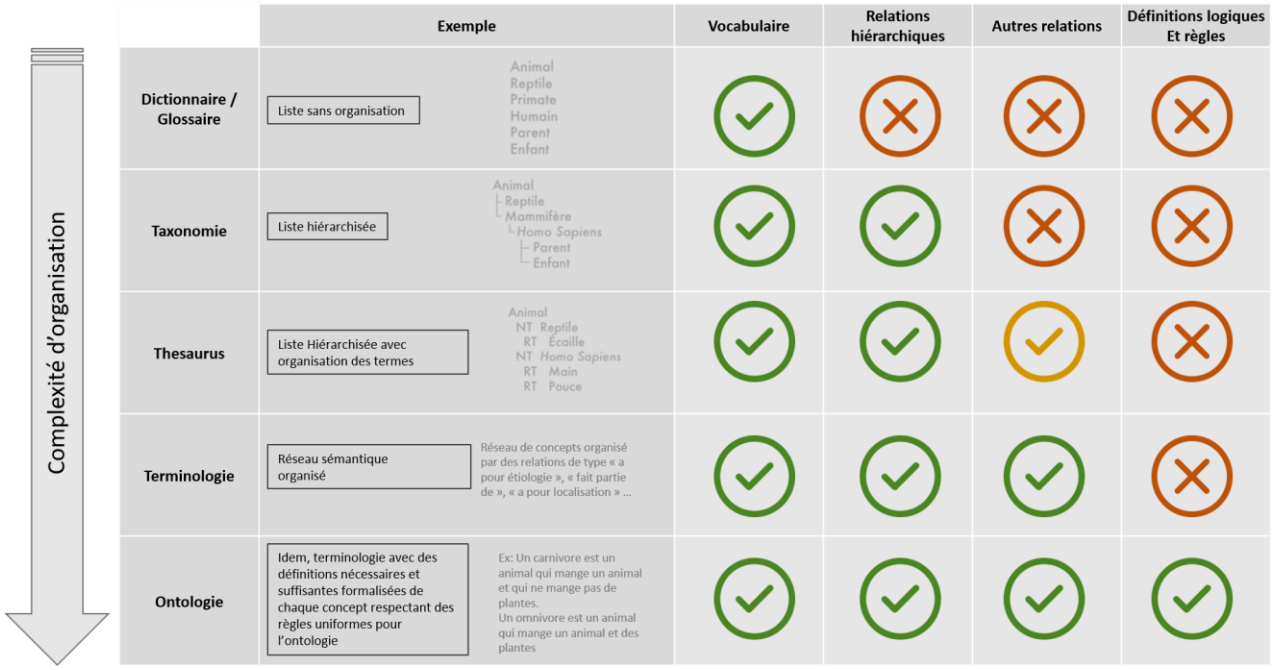
La figure ci-après présente les différentes ressources sémantiques disponibles en fonction de la complexité de leur organisation et de la richesse de leurs composants.

Une ressource sémantique est avant tout un vocabulaire pour décrire un domaine de connaissance. Ce vocabulaire peut être organisé en relations hiérarchiques (taxonomie, classification) auxquelles peuvent s'ajouter d'autres relations (synonymes, termes reliés) plus riches. On a alors affaire à des thesaurus ou des terminologies.

L'ontologie est un ensemble de concepts et de termes structuré en réseau représentant le sens d'un champ de connaissance. Elle peut être utilisée pour la transmission de connaissances, la recherche d'information et plus particulièrement la génération de nouvelles connaissances à partir de données existantes (déduction – si A = B et B = C, alors C = A – ou raisonnement – classification automatique de nouvelles connaissances qui respectent toutes les définitions déjà existantes).

Le formalisme des règles et leur uniformité distinguent la terminologie de l'ontologie.

Figure 9 : Classification des ressources sémantiques en fonction de l'organisation de leurs composants



3.3.1 Évaluation des ressources sémantiques

De nombreuses ressources sémantiques des domaines de la santé et du médico-social sont à disposition. Elles structurent plus ou moins formellement la connaissance, dans divers domaines médicaux.

Leur constitution passe par une étape de modélisation des connaissances médicales qui vise à définir des concepts et leur organisation. Chaque ressource sémantique possède ses propres caractéristiques : structure lexicale, relations entre concepts, règles logiques de construction de la ressource et formalisation des définitions des concepts.

Les cas d'usage des ressources sémantiques sont multiples, allant de l'interaction entre humains au raisonnement automatique (déductions et inférences) en passant par l'interopérabilité entre machines pour échanger des informations ou par l'indexation des documents médicaux pour en faciliter l'exploitation.

Il est donc important de pouvoir juger de la qualité d'une ressource sémantique mise en production sur des critères objectifs afin d'opérer le bon choix pour la communauté.

Avant de choisir la ou les ressource(s) sémantique(s) appropriée(s) pour chaque cas d'usage, il est donc essentiel de les évaluer en vérifiant que les concepts soient correctement représentés, d'une manière précise et cohérente tout en assurant une couverture optimale du domaine en question. Il est également crucial de vérifier si les ressources sont en adéquation avec les besoins des futurs utilisateurs.

Cette étude a permis de recenser les méthodes et critères d'évaluation pour les différents composants des ressources sémantiques, en fonction des cas d'usage.

L'évaluation de ressources sémantiques revêt un aspect multidimensionnel englobant l'évaluation des concepts, de leurs relations et des règles régissant leur structure, de leur maintenance et de leur adéquation à l'environnement d'intégration. **De nombreuses méthodes d'évaluation sont présentées dans la littérature, mais aucun consensus ne se dégage sur une approche globale systématique et acceptée par tous.**

Cependant, certains principes sont communément acceptés et permettent de guider les décisions vis-à-vis de l'adoption des ressources sémantiques. Le tableau ci-dessous propose une classification non-exhaustive permettant de prioriser les composants à évaluer en fonction des principales utilisations d'une ressource sémantique.

Figure 10 : Classification des composants à évaluer en fonction des cas d'usage

	Cas d'usage / application \ Composant évalué	Vocabulaire	Relations Hiérarchiques et Sémantiques	Définitions logiques et Règles	Composant contextuel
Interface Utilisateur	Interface Utilisateur; Encodage et codage manuel	X	X		Maintenance Consensus Satisfaction utilisateur Couplage Evolution et dérive Intégration organisationnelle Déployabilité Reconnaissance
Interface Utilisateur	Interopérabilité entre systèmes d'information Alignements de ressources sémantiques	X	X		
Dialogue machine à machine	Moteur de recherche et Indexation; Annotation et Extraction de connaissance	X	X	(X)	
Dialogue machine à machine	Raisonnement, Déduction et Inférence; Système d'aide à la décision; Intégration de base de données;	X	X	X	

Pour tous les cas d'usage, il est donc recommandé d'évaluer :

- Le composant « vocabulaire » pour vérifier notamment l'acceptation des termes par la communauté d'utilisateurs, l'adéquation par rapport à l'objectif d'usage et le niveau de couverture ;
- Le composant « Relations Hiérarchiques et Sémantiques » pour la précision (les connaissances conceptualisées dans la ressource sémantique sont-elles suffisamment précises au regard de l'expert ?) ou la cohérence des relations qui associent les concepts les uns avec les autres (la ressource sémantique contient-elle des contradictions ou des incohérences ?).

Pour les cas d'usage complexes requérant des ressources de type ontologique, il faut également évaluer les définitions logiques et les règles à partir desquelles les mécanismes de raisonnement automatique peuvent être effectués.

A côté de ces composants intrinsèques à la ressource sémantique, il existe des critères généraux liés à un composant environnemental comme les contraintes de maintenance, la satisfaction des utilisateurs, la reconnaissance par la communauté.

3.3.2 Statut de l'évaluation de la SNOMED CT

Dans le paysage des ressources sémantiques médicales, la SNOMED CT s'affirme comme étant la plus complète. Certaines propriétés empruntées aux modèles d'ontologies font de cette terminologie médicale multi-domaines une ressource sémantique hybride entre terminologie et ontologie.

La SNOMED CT est une ressource sémantique médicale qui bénéficie d'une communauté dédiée et de mises à jour régulières. Ce support technique permet donc l'assurance d'un suivi par des experts. Un autre avantage de la SNOMED CT est son niveau d'adoption par la communauté des systèmes d'information médicale à travers le monde. Son acceptation facilite d'autant plus l'échange entre les données au sein de systèmes d'information ayant adopté la SNOMED CT.

Au regard de l'interopérabilité avec d'autres ressources sémantiques, la SNOMED CT se positionne comme une terminologie pivot multi-domaines et peut s'interfacer avec les ressources sémantiques les plus utilisées (MedDRA, MeSH, LOINC, RxNorm, CIM-10-11)[30][31][32].

En revanche, la SNOMED CT présente des défauts d'assurance qualité au niveau de sa structure et de son vocabulaire. Particulièrement, une mauvaise définition dans les relations entre les concepts ou des relations manquantes peuvent impacter la qualité du raisonnement[33][34].

Par ailleurs, certains domaines de connaissance sont partiellement couverts par la SNOMED CT, comme montré lors de l'alignement avec Disease Ontology, notamment pour représenter les maladies rares et certaines morphologies tumorales. Cependant, la capacité d'alignement de la SNOMED CT avec des ressources sémantiques plus complètes[35][36] peut compenser ce manque de couverture.

La SNOMED CT est une ressource riche, mais qui reste complémentaire de terminologies de référence plus complètes dans leur domaine de connaissance dédié. La SNOMED CT peut servir de squelette autour duquel s'articuleraient des ressources sémantiques plus complètes ou adaptées aux cas d'usage envisagés.

La SNOMED CT n'étant pas une ontologie formelle dans son entier, son utilisation à des finalités de raisonnement doit se faire avec prudence. Les applications telles que les outils d'aide au diagnostic ou la génération d'hypothèses

[30] Kocbek S, Kim JD. Exploring biomedical ontology mappings with graph theory methods. PeerJ. 2017;5:e2990. Published 2017 Mar 2. doi:10.7717/peerj.2990.
[31] Rodrigues JM, Robinson D, Della Mea V, et al. Semantic Alignment between ICD-11 and SNOMED CT. Stud Health Technol Inform. 2015;216:790–794.
[32] Raje S, Bodenreider O. Interoperability of Disease Concepts in Clinical and Research Ontologies: Contrasting Coverage and Structure in the Disease Ontology and SNOMED CT. Stud Health Technol Inform. 2017;245:925–929.
[33] El-Sappagh S, Franda F, Ali F, Kwak KS. SNOMED CT standard ontology based on the ontology for general medical science. BMC Med Inform Decis Mak. 2018;18(1):76. Published 2018 Aug 31. doi:10.1186/s12911-018-0651-.
[34] Cui L. COHeRE: Cross-Ontology Hierarchical Relation Examination for Ontology Quality Assurance. AMIA Annu Symp Proc. 2015;2015:456–465. Published 2015 Nov 5.
[35] Héja G, Surján G, Varga P. Ontological analysis of SNOMED CT. BMC Med Inform Decis Mak. 2008;8 Suppl 1(Suppl 1):S8. Published 2008 Oct 27. doi:10.1186/1472-6947-8-S1-S8.
[36] Halland K, Britz K, Gerber A. Investigations into the use of Snomed CT to enhance an OpenMRS health information system. South African Computer Journal. 2011;Vol. 47. Published 2008 Oct 27. doi: 10.18489/sacj.v47i0.14

par algorithme ne devraient pas uniquement reposer sur les propriétés ontologiques de la SNOMED CT, mais aussi intégrer des contrôles internes permettant de s'assurer de leur fiabilité.

3.3.3 Résumé : ce que ce chantier nous apprend sur l'évaluation terminologique et l'évaluation de la SNOMED CT dans la littérature

Tableau 3 : Chantier bibliographique

Avantages	Inconvénients
- Il existe de nombreuses méthodes pour évaluer les Terminologies. - La SNOMED CT est une ressource sémantique riche, multi-domaines et ayant un positionnement de terminologie « pivot ». - La SNOMED CT est une ressource sémantique évolutive bénéficiant d'une communauté médicale internationale et d'une maintenance régulière.	- Aucun consensus ne se dégage sur une approche d'évaluation sémantique, globale systématique et acceptée par tous. - La SNOMED CT a une couverture partielle de différents domaines de connaissance et demeure complémentaire d'autres ressources de « référence » (Oncologie, maladies rares, anatomie) - Des différences éditoriales par rapport à des terminologies de référence (ex : CIM-11) imposent un processus complexe d'alignement. - Un usage limité en matière de raisonnement.

3.4 Analyse du positionnement ou des performances de la SNOMED CT dans différents cas d'usage

Cette partie présente les résultats des études liées au chantier scientifique.

Dans le cadre des études ci-après, relatives à divers cas d'usage ou domaines du champ santé-social, l'objectif pour plusieurs a été de mesurer la couverture respective des domaines étudiés par certaines Terminologies. De la notion de couverture découle le principe de complétude terminologique (« *completeness* » en anglais).

Une Terminologie peut être considérée comme « complète » si elle couvre l'ensemble des concepts du domaine qu'elle cherche à représenter. Une Terminologie est vivante, et ne peut être « complète » qu'à un instant « t ». Une Terminologie « complète » n'est donc pas pour autant « terminée ». La complétude revient à permettre aux professionnels de santé auxquels la Terminologie s'adresse de manipuler tous les éléments nécessaires à l'exercice de leur métier au travers de leurs logiciels métiers.

A défaut de pouvoir recenser tous les cas d'usages santé-social et les couvertures respectives de ces cas d'usages par les Terminologies existantes, les entrepôts de données de santé (EDS) hospitaliers concentrent les données de soins produites par les systèmes informatiques utilisés dans le parcours hospitalier : les données des EDS sont un « concentré » des cas d'usage existants. L'annotation systématique de cette masse de données en utilisant les concepts portés par les Terminologies permet de vérifier si elles modélisent, de manière pertinente, les concepts contenus dans les dossiers médicaux français.

3.4.1 Cas d'usage microbiologie (AP-HP/ANS)

Les détails de l'étude sont présentés en **annexe P4.1**

3.4.1.1 Contexte et méthode

L'AP-HP a intégré, au sein de son Entrepôt de Données de Santé, les résultats des analyses microbiologiques réalisées lors de la prise en charge des patients au sein de ses établissements. Elle souhaite standardiser cette base de données et la connecter avec des bases de données dans d'autres hôpitaux.

Dans cet objectif, l'AP-HP et les hôpitaux partenaires souhaitent adopter une terminologie de référence pour traiter leur besoin d'échange de données microbiologiques.

Le présent travail s'inscrit dans une démarche évaluative pour rationaliser la préconisation d'une terminologie du domaine microbiologique. A noter que ce travail s'inscrit également dans la démarche du COPIL CR de biologie animé par l'ANS qui a décidé de rechercher une ou des terminologies de microbiologie dans le cadre du décret biologie de 2016.

Cette étude répond au double objectif :

- comparer les performances relatives des terminologies du domaine de la microbiologie utilisables dans les cas d'usage de surveillance sanitaire et de recherche épidémiologique ;
- recommander une terminologie de référence du domaine.

L'AP-HP possède un catalogue de germes recensant 10 146 bactéries, qu'elle souhaiterait codifier précisément.

La recherche d'une terminologie de référence est menée en 3 étapes :

1) Identification des terminologies de référence dans le domaine de la microbiologie sur le serveur multi-terminologie du National Center of Biomedical Ontologies (NCBO Bioportal) sur la base d'un échantillon de 15 bactéries issues du catalogue de l'AP-HP ;
2) Evaluation des terminologies candidates sur des critères liés au vocabulaire (couverture, précision), aux relations entre concepts ainsi que sur des critères d'accessibilité et d'exploitabilité ;
3) Analyse de l'alignement des terminologies candidates par rapport à la base complète de bactéries de l'AP-HP.

3.4.1.2 Résultats

- Identification des ressources sémantiques candidates

Tout ou partie de l'échantillon du catalogue de l'AP-HP se retrouve dans 24 terminologies présentes sur Bioportal.

Tous les genres bactériens sont retrouvés dans la SNOMED CT et dans NCBI Taxonomy. Les 15 espèces sont présentes dans NCBI Taxonomy et 13/15 dans la SNOMED CT.

A noter qu'une grande partie des bactéries de l'échantillon sont également retrouvées dans la LOINC (10/15), NCIT (National Cancer Institute thesaurus) (10/15), la CIM-11 (11/15) et dans le thésaurus MeSH (11/15).

NCBI Taxonomy et SNOMED CT sont les deux ressources sémantiques les plus appropriées pour coder des bactéries.

- Evaluation de l'accessibilité et de l'exploitabilité des vocabulaires

Le tableau ci-après synthétise les principales observations.

Tableau 4 : Evaluation de l'accessibilité et de l'exploitabilité des vocabulaires

Critère	SNOMED CT	NCBI Taxonomy	Conclusions
Périmètre de la terminologie	► 6 391 bactéries codées au niveau genre espèce ► (11 508 bactéries avec des nommages hybrides taxonomie / groupes)	► 371 766 bactéries définies au niveau genre espèce (taxonomie pure)	► **Couverture supérieure de NCBI Taxonomy** par rapport à la SNOMED CT ► Couverture insuffisante de SNOMED CT par rapport au catalogue de l'AP-HP (10146 bactéries)
Nommage des bactéries	► Nom principal ► **Peu de synonymes** ► Nommage spécifique à un cas d'usage (coagulase negative Staphylococcus species, not Staphylococcus lugdunensis, Enterohemorrhagic Escherichia coli, serotype O50:H7, Salmonella II 50:z10:z6:z42)	► Nom principal ► Présence de **nombreux synonymes**	► Meilleure précision de NCBI Taxon vs SNOMED CT au niveau de synonymes ► Le manque de synonymes dans la SNOMED CT peut entraîner des faux négatifs en annotation. ► Le détail des groupes bactériens au sein de la SNOMED CT entraîne un encombrement de concepts précoordonnés qui pourrait être traité en post coordination (ex : La SNOMED CT dénombre 2707 salmonella enterica ,667 escherichia coli). Ceci explique la différence entre les 6391 bactéries définies au niveau genre espèce par rapport aux 11508 concepts du domaine bacteria
Classification	► Multiaxiale – taxonomique et morpho-biochimique (ex : propriété enzymatiques, formes, et propriété des parois bactériennes…)	► Mono-axiale taxonomique pure	► Une richesse de classification plus grande pour la SNOMED CT, mais fluctuante. ► NCBI Taxonomy présente une arborescence taxonomique stricte et uniforme
Données liées	► Non (il n'existe pas de liens vers des ressources externes sur le navigateur SNOMED CT)	► Oui (il existe de nombreux accès à de multiples ressources NCBI génomiques et à des ressources externes via l'identifiant taxon)	► L'ouverture des données est meilleure avec NCBI Taxonomy qu'avec la SNOMED CT

Tableau 5 : Evaluation de l'accessibilité et de l'exploitabilité des vocabulaires (suite)

Critère	SNOMED CT	NCBI Taxonomy	Conclusions
Prix Mise à jour Source	► La gratuité pour les utilisateurs n'est assurée que par le paiement d'une licence nationale par l'état. Dans le cas contraire, paiement par les utilisateurs de licences dites « affiliées ». ► Mise à jour par release de 6 mois ► Source privée	► Licence gratuite ► Mise à jour quotidienne ► Source publique	► NCBI Taxonomy est compatible avec la politique d'ouverture des données de l'état français

- Evaluation de l'alignement du catalogue AP-HP avec NCBI Taxonomy et SNOMED CT

Le tableau ci-après synthétise les principales observations.

Tableau 6 : Evaluation de l'alignement du catalogue AP-HP avec NCBI Taxonomy et SNOMED CT

Critère	SNOMED CT	NCBI Taxonomy	Conclusion
►Alignement de la terminologie (par rapport au catalogue AP-HP)	**58.2%** **Exact match**	**79.2%** **Exact match**	►NCBI Taxonomy présente une meilleure qualité d'alignement que SNOMED CT
►Exemple qualitatif ►genre Helicobacter (33 espèces dans le catalogue AP-HP)	24 espèces ►Couverture : **73%**	►59 espèces + 161 en cours d'identification ►Couverture : **>179%** *(NCBI Taxonomy va au-delà du catalogue de l'AP-HP)*	►NCBI Taxonomy présente une meilleure couverture que SNOMED CT sur l'exemple choisi

3.4.1.3 Conclusion

Cette étude a permis d'identifier SNOMED CT et NCBI Taxonomy comme terminologies de référence dans le domaine de la microbiologie.

En termes d'évaluation des vocabulaires, NCBI Taxonomy a une couverture et une précision supérieure par rapport à la SNOMED CT.

SNOMED CT présente une couverture insuffisante par rapport au catalogue de bactéries de l'AP-HP.

En termes de relations hiérarchiques, la SNOMED CT intègre des arbres taxonomiques et des arbres de typage morpho-biochimiques qui permettent une richesse de classification, mais qui peuvent fluctuer. NCBI Taxonomy présente une arborescence taxonomique stricte et uniforme.

Du point de vue droit de propriété intellectuelle, NCBI Taxonomy est dans le domaine public alors que SNOMED CT est publiée sous licence propriétaire.

Ces observations positionnent NCBI Taxonomy comme une meilleure candidate pour constituer un vocabulaire d'interopérabilité du domaine microbiologique par rapport à SNOMED CT.

Sur la base de l'évaluation réalisée à l'AP-HP, NCBI Taxonomy est la meilleure candidate pour standardiser un catalogue hospitalier de souches bactériennes permettant l'interopérabilité et l'exploitation des bases de données d'analyses microbiologiques et faire des passerelles vers d'autres sources de connaissance.

A l'issue de cette étude, l'AP-HP s'est engagée dans la finalisation de l'alignement de son catalogue avec NCBI Taxonomy et a initié des échanges avec l'Institut Pasteur et des hôpitaux partenaires afin d'évaluer la faisabilité de l'adoption de cette terminologie.

3.4.2 Cas d'usage anatomie (LIMICS)

Les détails de l'étude sont présentés en **annexe P4.2**

3.4.2.1 Contexte et méthode

L'anatomie est un domaine de connaissance fondamental utilisé dans de multiples cas d'usages de santé : localisation des atteintes, des actes médicaux, imagerie, tissuthèque, etc.

Une terminologie de référence est nécessaire pour décrire ce domaine de connaissance avec une couverture et une précision pertinente pour l'utilisateur.

La présente étude s'est focalisée spécifiquement sur 5 terminologies couvrant le domaine de l'anatomie :

- **SNOMED CT** : La SNOMED CT a été étudiée après souscription de licence de licence affiliée. Elle n'est pas accessible publiquement ;
- **CIM-11** : On accède à la CIM-11 via l'URL : https://icd.who.int/dev11/f/en#/ On accède à une linéarisation de la CIM-11 à l'URL : https://icd.who.int/browse11/l-m/en ; La CIM-11 est téléchargeable par les APIs de l'OMS[37] ;
- **FMA** : URL : http://si.washington.edu/content/comparisons-other-anatomy-sources ;
- **UBERON** : URL : www.Uberon.org ;
- **MeSH** : URL : https://www.nlm.nih.gov/mesh/meshhome.html.

La SNOMED CT et la CIM-11 ont été comparées aux deux terminologies de référence dans le domaine de l'anatomie (Uberon et FMA) auxquelles ont été ajoutées les terminologies de référence pour indexer les documents médicaux.

Il est à noter que de nombreuses autres ressources sémantiques incluent des descriptions anatomiques telles que le National Cancer Institute Thesaurus (NCIT) ou la classification d'anatomo-pathologie française ADICAP par exemple.

L'objectif de ce livrable est d'étudier les cinq RTO ou SOC sélectionnés, tous intéressants vis-à-vis du cas d'usage anatomie.

Cette étude comprendra le positionnement ces cinq RTO les unes par rapport aux autres en termes de :

- analyse quantitative des concepts présents en anglais et en français en général, puis ceux spécifiques en anatomie ;
- alignements inter-terminologiques sur ces cinq RTO, impliquant la couverture des différentes RTO entre elles ;
- couverture terminologique respective de ces cinq RTO par rapport à des millions de documents de santé d'un entrepôt du CHU de Rouen ; cet élément permettra la capacité d'annotation de données de santé ;
- limites, avantages et inconvénients de chaque terminologie seront discutés, ce qui permettra d'évaluer leurs performances relatives.

3.4.2.2 Résultats

Cette analyse présente comme résultats principaux :

- En termes de couverture, la **FMA présente un nombre de concepts spécifiques au cas d'usage anatomie plus important que la SNOMED CT** : 105 072 vs 29 636 ou la CIM-11 (3626). **La FMA offre ainsi la possibilité d'une couverture quasiment exhaustive du domaine anatomie** ;

[37] https://icd.who.int/icdapi

Tableau 7 : Nombre de concepts et de traductions en français des terminologies d'anatomie

Terminologie	Nombre total de concepts	Nombre de concepts traduits en français	% de concepts traduits en français	Nombre de concepts en anatomie	% de concepts en anatomie	Nombre de concepts traduits en français	% de concepts traduits en français
FMA*	105072	17158	16,33%	105072	100,00%	17158	16,33%
Uberon**	15141	10	0,07%	13776	90,98%	10	0,07%
CIM-11***	59709	55278	92,58%	3626	6,07%	3293	90,82%
SNOMED CT****	326947	202292	61,87%	29636	9,06%	15120	51,02%
MeSH	378 087	106 000	28,04%	4166	1,10%	4166	100%

** FMA : tous les concepts*
*** Uberon : concepts fils de 'Anatomical entity' 0001062*
**** CIM-11 : concepts fils de 'Anatomy and topography' 1154280071*
***** SNOMED CT : concepts fils de 'Anatomical structure' 91723000 + type sémantique UMLS fils de 'Anatomy'*

- En termes d'alignements, en considérant ceux manuellement validés, c'est la FMA qui s'aligne le plus avec chacune des quatre autres terminologies. **C'est le résultat le plus important de ce travail, qui confirme les données de la littérature ou avis d'experts terminologiques qui considèrent FMA comme la RTO de référence pour l'anatomie.**
- En termes de capacité d'annotation, la FMA présente une **meilleure capacité d'annotation** (n=6522 termes annotants)**.** Elle est suivie par la SNOMED CT (n=4503), le MeSH (n=2689) et la CIM-11 (n=1676).
- **Ainsi de façon attendue, la FMA qui est objectivement l'ontologie de référence en anatomie, confirme son statut avec le nombre le plus important de concepts uniques annotants sans lien avec d'autres terminologies (n=4394), suivi par la CIM-11 (n=1435) et la SNOMED CT (n=1394). Sur ce critère, la CIM-11 et la SNOMED CT font jeu égal. Ceci montre leur complémentarité en annotation.**
- En comparaison, la SNOMED CT présente un plus grand nombre de concepts non-opérationnels (concepts non-retrouvés) pour ce cas d'usage : 70,34 % des termes de la SNOMED CT n'ont pas été trouvés au sein de l'entrepôt de données du CHU de Rouen contre 63,83 % pour la FMA. Le MeSH et la CIM-11 sont bien positionnés sur ce critère d'adéquation aux besoins avec respectivement 35.12% et 49.12% de concepts non-opérationnels seulement.
- Une comparaison de la SNOMED CT et de la CIM-11 permet de conclure que, dans l'attente d'une étude qualitative des concepts annotants trouvés, la SNOMED CT est supérieure à la CIM-11 sur sa capacité d'annotation en anatomie (4 503 vs 1 676).

Tableau 8 : Capacité d'annotation de CIM-11, FMA, SNOMED CT, Uberon et MeSH sur les documents de l''entrepôt EDSaN (CHU Rouen)

Terminologies	Total	Nb concepts trouvés	Nb concepts non trouvés	Nb concepts uniques sans lien avec les autres terminologies
CIM-11	3294	1676 (50,88%)	1618 (49,12%)	1435 (43,56%)
FMA	17812	6522 (36,62%)	11290 (63,38%)	**4394** (24,67%)
SNOMED CT	15180	4503 (29,66%)	10677 (70,34%)	1394 (9,18%)
Uberon	14	4 (28,57%)	10 (71,43%)	2 (14,29%)
MeSH	4166	2689 (64,55%)	1477 (35,45%)	761 (18,27%)

3.4.2.3 Limites de l'étude

- Les travaux d'annotation ont été réalisés sur un seul Entrepôt de données (EDS). Les résultats seraient à confirmer sur un autre EDS ;
- Manque d'approfondissement de Uberon terminologie de référence en anatomie en raison de l'absence de traduction française ;

- Traduction partielle de FMA sous estimant sa capacité d'annotation ;
- Le travail a été essentiellement quantitatif. Il pourra être complété par une approche qualitative pour renforcer l'argumentation des forces et faiblesses de chaque ressource sémantique et étudier l'adéquation des vocabulaires par rapport aux besoins des utilisateurs professionnels de santé.

3.4.2.4 Conclusion

Globalement, la FMA est la terminologie de référence du point de vue quantitatif sur l'ensemble des critères étudiés dans ce travail : (a) en nombre absolu de concepts en anglais et français couvrant l'anatomie (plus en anglais qu'en français) ; (b) c'est l'ontologie qui est la plus alignée avec les quatre autres ; (c) c'est l'ontologie qui fournit le plus de concepts retrouvés dans des documents de santé, à la fois dans l'absolu et en regardant plus précisément les concepts qui ne sont pas couverts par les autres, autrement dit pas alignés avec les autres RTO à l'étude.

Dans l'attente d'une étude qualitative des concepts annotants trouvés, la SNOMED CT est supérieure à la CIM-11 sur sa capacité d'annotation en anatomie.

Cependant, la SNOMED CT et la CIM-11 sont similaires sur le plan des concepts uniques (singularités) d'annotation, avec un avantage pour la CIM-11 sur le pourcentage des concepts uniques.

La SNOMED CT sur-modélise l'anatomie : 70% des concepts sont non-annotants (non-opérationnels) et appelleraient donc des efforts de maintenance et de traduction superflus.

3.4.3 Cas d'usage maladie de Charcot (ou SLA) (LIMICS)

Le détail de cette étude est présenté en **annexe P4.3.**

3.4.3.1 Contexte et méthode

Les maladies rares en France représentent un énorme enjeu en termes de coordinations de soins et de suivi du patient.

En France, elles représentent un enjeu majeur de santé publique. Si chacune des maladies ne touche qu'un faible nombre de patients, il existe plus de 7 000 maladies rares identifiées à ce jour. La prévalence globale atteint ainsi plus de 3 millions de personnes soit 4,5% de la population.

La présente étude se focalise sur les patients atteints de Sclérose Latérale Amyotrophique (SLA ou maladie de Charcot). Elle s'inscrit dans l'optimisation du parcours de soins des patients avec fluidification des échanges par une meilleure fluidité des échanges interprofessionnels.

Les patients atteints de SLA nécessitent un accompagnement pluridisciplinaire au cours de leur parcours. La pathologie provoque de nombreuses incapacités et situations de handicaps. La prise en charge de cette maladie est complexe peut amener à des situations de rupture de parcours par l'absence, l'arrêt ou des difficultés de prise en charge.

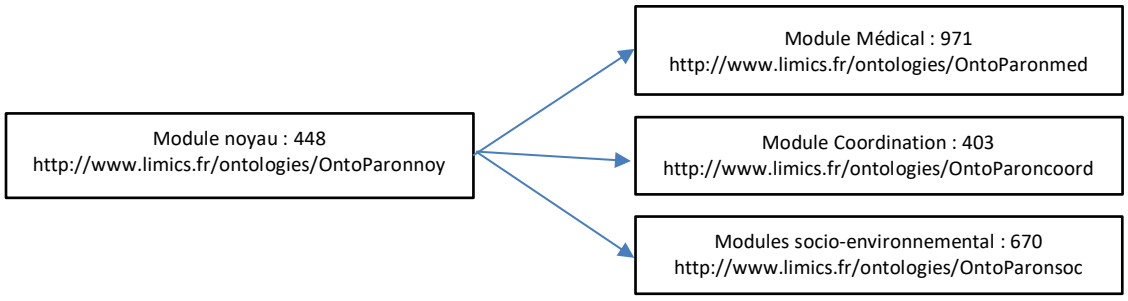
Une ontologie de domaine OntoParon a été créée par le LIMICS afin d'analyser les parcours de soins des patients atteints de SLA (besoins/demandes patients, ruptures de parcours)[38].

Cette étude compare les terminologies disponibles sur heTOP[39] (contenant CIM-11 et SNOMED CT) à OntoParon en termes de couverture.

L'objectif est de voir si une ou plusieurs terminologies existantes peuvent se substituer à OntoParon pour analyser le parcours de soins du patient SLA.

L'ontologie OntoParon développée dans le cas de la Sclérose Latérale Amyotrophique (SLA) est une ressource permettant l'annotation sémantique des données textuelles d'un réseau de coordination, le réseau SLA Île-de-France. OntoParon est une ontologie modulaire permettant de prendre en compte plusieurs domaines du parcours de soins. Le schéma ci-dessous illustre les différents modules.

Figure 11 : Représentation schématique d'OntoParon (nombre de classe/modules)



- le module noyau de 448 classes, contient l'ensemble des concepts de haut niveau communs aux trois ontologies, comme les objets idéaux, les agents, les processus, les modalités ;

[38] From Sonia Cardoso - https://hal.sorbonne-universite.fr/tel-02429414v1
[39] HeTOP est le serveur de terminologie interlingue du DI2M (Département d'Informatique et d'Information Médicales du CHU de Rouen) : https://www.hetop.eu/hetop/

- un module de coordination de 403 classes, contient les missions spécifiques de coordination (actions de communication, actions d'évaluation des besoins, actions de recherche de ressource (professionnel de santé, structure d'accueil) ;
- un module médical de 971 classes, spécifique du domaine de la médecine, (agents médicaux (médecin, neurologue, kinésithérapeute, etc.), processus médicaux (consultation, hospitalisations, etc.) et objets médicaux (les médicaments, les prescriptions, etc.). Dans ce module se trouve également, les concepts liés aux structures anatomiques, les signes et les symptômes. Il s'agit de l'ensemble des concepts liés directement à la pathologie et à la prise en charge médicale ;
- un module socio-environnemental de 670 classes, regroupe l'ensemble des concepts liés à la vie de la personne dans son environnement familial et social (situation professionnelle, situation sociale, mode de vie, types d'allocations sociales perçues). Y sont modélisés les agents sociaux (la famille, les auxiliaires de vie, etc.), les actions sociales (la demande de prestation de compensation du handicap, la demande d'aide humaine, etc.), les objets physiques du domaine social (fauteuil roulant, carte vitale, le logement, etc.), les concepts liés à la protection juridique de la personne (la tutelle ou curatelle).

Des alignements ont été réalisés entre OntoParon et les terminologies de HeTOP en intégrant OntoParon (termes préférés et synonymes en français) dans le serveur de terminologies. Les alignements avec les 85 terminologies de HeTOP ayant intégré la SNOMED CT et la CIM-11 ont ainsi pu être mesurés.

3.4.3.2 Résultats

Cinquante et une des 85 terminologies de heTOP ont au moins un alignement avec OntoParon. Au total, 9906 alignements sont trouvés. (6691 en excluant les doublons). Ces résultats permettent d'identifier 3 catégories de RTO en termes de pourcentages d'alignements obtenus. Il est à noter que le pourcentage d'alignement maximal est de 32.05% (pour le MeSH) :

- 4 RTO s'alignent à plus de 25% avec OntoParon (MeSH, SNOMED CT, NCIT, Thesaurus de santé publique)
- 4 RTO s'alignent à plus entre 10 et 15% avec OntoParon (ICNP, CIM11, MedDRA, LOINC) ;
- 12 RTO s'alignent à moins de 10% avec OntoParon (Radlex, WHO-ART, CISP2, Base médicament, HPO, CisMef, ICF, CIM-10, FMA, Cladimed, Nomenclature Orphanet)).

Les autres terminologies s'alignent marginalement (<1.2%).

Le tableau ci-après indique le niveau de couverture obtenu par module. Les sept premières RTO qui présentent le meilleur taux d'alignement y sont incluses.

Tableau 9 : Niveau de couverture par module

Terminologies	Total	Nombre de concepts trouvés sans doublon	Module Socio-environnemental : 670 classes (% Alignement)	Module médical : 971 classes (% alignement)	Module Coordination : 403 classes (% alignement)
Medical subject headings (MeSH)	277 575	789	181(27,01%)	501(51,6%)	14(3,47%)
SNOMED CT	350 976	763	163 (24,32%)	432(44,5%)	14(3,47%)
National Cancer Institute Thesaurus (NCIt)	79 870	698	127 (18,95%)	370(38,10%)	15(3,72%)
Thésaurus santé Publique (TSP)		653	195 (29,10 %)	348(35,8%)	13(3,22%)
Medical Dictionary for Regulatory Activities (MedDRA)	68 140	283	22 (3,28 %)	59(6,07%)	0(0%)
International classification for nursing Pratice (ICNP)	4 477	346	81 (12,08%)	165(16,99%)	9(2,23%)
CIM-11	55 267	310	39 (5,82 %)	243(25,02%)	3(0,74%)

Il est à noter que pour les trois modules d'OntoParon, les meilleurs taux d'alignements de classes obtenus étaient lors des analyses comparatives avec les quatre RTO suivantes : MeSH, SNOMED CT, NCIT et TSP.

- Module médical (dit PARON_MED) : Il s'agit du module d'OntoParon pour lequel le pourcentage d'alignements obtenu est le plus important. Concernant les RTO comparées au module médical, MeSH s'aligne à 51.6%, SNOMED CT à 44.5%, NCIT à 38.10% et TSP à 35.8%. Il est à noter que la CIM-11 s'aligne à 25.02% avec le module médical d'OntoParon ;
- Module socio-environnemental (dit PARON_SOC) : Concernant les RTO comparées au module socio-environnemental, TSP s'aligne à 29.10%, MeSH à 27.01%, SNOMED CT à 24.32% et NCIT à 18/95% ;
- Module de coordination (dit PARON_COORD) : Il s'agit du module d'OntoParon pour lequel le pourcentage d'alignements obtenu est le plus faible. Concernant les RTO comparées au module de coordination, NCIT s'aligne à 3.72%, MeSH et SNOMED CT toutes les deux à 3.47% et TSP à 3.22%.

Les résultats quantitatifs montrent que les alignements sont retrouvés principalement avec quatre RTO, cependant le pourcentage d'alignement n'excède de manière globale jamais plus de 33%. Au regard des résultats obtenus, l'utilisation seule des RTO existantes pour réaliser l'annotation des données de parcours de santé des patients ayant une SLA n'aurait pas permis d'annoter et d'extraire l'ensemble des informations recherché permettant la description des parcours de santé.

3.4.3.3 Conclusion

Ce travail permet de mettre en évidence qu'une seule RTO ne suffit pas à couvrir les classes créées dans le cadre de l'étude de parcours de santé dans la sclérose latérale amyotrophique (SLA). Il y a nécessité d'avoir plusieurs ressources sémantiques complémentaires pour couvrir un parcours de soins complexes (médical, médico-social, coordination). Le domaine de la coordination est pour l'instant le plus difficile à couvrir avec les ressources existantes.

La **SNOMED CT** et la **CIM-11** couvrent respectivement 30% et 12,59% des concepts nécessaires à décrire le parcours patient (référence ONTOPARON construite Ad hoc), démontrant leur **insuffisance marquée** dans le module de coordination. Le module médical est le mieux couvert.

Dans le cas où il n'y a pas d'ontologie de domaine, le travail mené met en évidence l'importance d'utiliser un serveur multi-terminologique avec des RTO alignées, afin de repérer un maximum de concepts. L'utilisation d'une seule ressource comme la SNOMED CT, ne permet pas de couvrir tous les domaines dans un cas d'usage spécifique comme les parcours de santé des patients ayant une SLA. Dans notre cas, la SNOMED CT ne couvre que 30,9% des concepts nécessaires.

3.4.4 Cas d'usage oncologie (ISPED)

Le détail de cette étude est présenté en **annexe P4.4.**

3.4.4.1 Contexte et méthode

Ce travail, réalisé par l'équipe ERIAS/ISPED, vise à étudier la couverture des connaissances liées à l'oncologie par les ressources sémantiques de référence suivantes : la SNOMED CT, la CIM-11 et la CIM-O3. Le choix s'est porté sur ces ressources, car la CIM-O3 est la ressource sémantique de référence utilisée par les registres de cancer, la SNOMED CT est la ressource sémantique la plus large et la plus utilisée dans le domaine biomédical et la CIM-11 est la nouvelle version de la CIM que l'Organisation Mondiale de la Santé (OMS) préconise d'adopter à partir de 2022 à la place de la CIM-10.

Deux approches distinctes ont été suivies. Tout d'abord, les concepts morphologiques et topographiques décrits dans la CIM-O3, la CIM-11 et la SNOMED CT ont été comparés d'après leurs termes normalisés en français et en anglais, ainsi que les concepts concernant les néoplasmes présents dans la CIM-11 et la SNOMED CT (Tableau 6). Cette méthode vise à évaluer la complétude du domaine de connaissances de l'oncologie par ces ressources sémantiques. Ensuite, l'adéquation des termes décrits dans ces trois ressources sémantiques avec le langage utilisé par les professionnels dans les documents textuels du CHU de Bordeaux a été étudiée. Cette méthode apporte des éléments de discussion concernant les critères de précision et de véracité, en plus de la complétude qui est également considérée.

3.4.4.2 Résultats

Tableau 10 : Taux de couverture des concepts de morphologie, topographie et diagnostic (néoplasmes) entre la CIM-O3, la CIM-11 et la SNOMED CT (d'après les termes en anglais et en français). La CIM-O3 ne décrivant pas de diagnostics, seule la couverture entre la CIM

Axe	Langue des termes	Couverture de la CIM-O3 par la CIM-11	Couverture de la CIM-O3 par la SNOMED CT	Couverture de la CIM-11 par la SNOMED CT
Morphologie	anglais	92,4%	90,4 %	75,9 %
	français	39 ,0%	33,0 %	37,6 %
Topographie	anglais	57,5%	62,8%	51,3%
	français	29,1%	19,1 %	27,0%
Diagnostic	anglais	/	/	27,2 %
	français	/	/	16,1 %

La première approche a permis de montrer que la couverture des libellés de concepts de morphologie de la CIM-O3 par la SNOMED CT et la CIM-11 était très bonne (plus de 90%). Ces résultats étaient attendus puisque ces deux dernières ressources sémantiques ont utilisé la CIM-O3 comme support pour la description de leurs concepts morphologiques.

En ce qui concerne la topographie, les résultats obtenus sont plus mitigés avec une couverture de la CIM-O3 par la CIM-11 et la SNOMED CT autour de 60%. L'existence de termes complexes et de noms de catégories résiduelles (termes contenant des mots tels que « other ») dans la CIM-O3 peut expliquer ces différences avec la SNOMED CT. Nous n'avons néanmoins pas compris la raison pour laquelle la couverture n'était pas plus élevée par la CIM-11 qui contient pourtant ce type de concepts.

Finalement, la couverture des concepts de néoplasmes décrits dans la CIM-11 par la SNOMED CT est relativement basse (moins de 30%). Comme les autres CIM, la CIM-11 contient des concepts décrivant des catégories résiduelles nécessaires pour le codage de l'ensemble des diagnostics observés chez les patients. Les concepteurs de la SNOMED CT ont fait le choix de décrire les concepts biomédicaux de manière exhaustive et donc de ne pas représenter de tels concepts.

Deux autres enseignements ont été tirés de cette étude de couverture d'après les libellés de concepts. D'une part, les traductions en français de la CIM-11 et de la SNOMED CT qui ont été générées automatiquement sont prometteuses, car les versions françaises de ces ressources sémantiques ont permis d'identifier des appariements avec la version française de la CIM-O3 qui est mise à disposition par l'OMS (et dont les libellés en français sont donc validés). Cependant, la méthode de traduction automatique nécessite d'être améliorée, car un certain nombre de problèmes ont été rencontrés lors de l'analyse des résultats (ordre des mots inapproprié, présence d'articles définis inutiles, mots anglais non traduits, etc.). D'autre part, l'analyse morphologique utilisée dans cette étude pour comparer les libellés pourrait être enrichie, notamment en utilisant des synonymes connus, afin d'obtenir une meilleure couverture entre la SNOMED CT et la CIM-11.

Notons enfin que chacune des trois ressources sémantiques considérées décrit des concepts qui lui sont propres en plus des concepts communs identifiés dans ce travail. **Il est donc légitime de conclure qu'une meilleure couverture du domaine de l'oncologie serait obtenue en utilisant l'ensemble de ces ressources qui apparaissent comme complémentaires (observation qui concerne plus particulièrement la SNOMED CT et la CIM-11).**

En ce qui concerne l'adéquation des termes, d'une manière générale, la SNOMED CT permet d'indexer plus de documents sur les trois axes (topographie, morphologie et diagnostic) que la CIM-O3 et la CIM-11. Cependant, il est à noter qu'un certain nombre de termes inclus dans ces axes de la SNOMED CT sont peu informatifs (très ambigus ou non-adaptés). A titre d'exemple, les termes « interne » ou « tout » sont retrouvés respectivement dans les axes de topographie et de morphologie de la SNOMEDCT et contribuent à indexer de nombreux documents. Ce constat peut s'expliquer par des erreurs de traduction ou par l'inclusion de concepts non adaptés lors de la sélection des axes. Par ailleurs, la version de la CIM-O3 mise à disposition ne comprend pas les termes correspondant aux inclusions, ce qui limite le vocabulaire disponible pour cette ressource. Un constat important est que la CIM-11 ainsi que la SNOMED CT, de par leur structuration, permettent d'apporter des termes génériques (tumeur, cancer, tumeur du pancréas, etc.) qui sont utilisés de façon fréquente dans les documents**. De ce point de vue, ces ressources présentent un vocabulaire riche et structuré. Enfin, il a été constaté une complémentarité des ressources évaluées sur les différents axes et aucune ne permet d'indexer l'ensemble des termes à elle seule.**

Tableau 11 : Nombre de termes et de documents indexés suivant les différents axes terminologiques et nombre de concepts distincts correspondants

Ressource sémantique	Nombre de termes	Nombre de concepts	Nombre de documents
Topographie			
CIM-O3	340 777	688	49 199
CIM-11	282 448	846	49 047
SNOMED CT	432 102	1 311	70 312
Morphologie			
CIM-O3	51 015	311	28 318
CIM-11	57 743	368	31 911
SNOMED CT	73 409	283	36 643
Diagnostic			
CIM-11	32 488	221	15 888
SNOMED CT	72 344	329	33 168

3.4.4.3 Conclusion

Par rapport à la CIM-O3 de référence, **SNOMED CT et CIM-11 ont une couverture équivalente** en morphologie (> 90%) et en topographie (entre 58 et 62%).

CIM-11 et SNOMED CT ne sont pas équivalentes dans la description des néoplasmes du fait de différences éditoriales. Les travaux d'alignements doivent être approfondis en prenant en compte les différences

Il y a au sein de la SNOMED CT la présence de singularités comme dans la CIM-11, suggérant les complémentarités entre les deux ressources pour obtenir la meilleure couverture du domaine oncologie.

Il semble nécessaire de mettre à disposition l'ensemble de ces trois ressources et de les lier entre elles afin de couvrir le vocabulaire utilisé en routine et ainsi favoriser une indexation large et efficace de l'information produite lors du soin.

Ce travail s'est focalisé sur le composant « vocabulaire » des ressources sémantiques et les deux cadres d'analyse ont permis de conclure que l'usage concomitant de la SNOMED CT et de la CIM-11 est recommandé pour garantir une meilleure couverture du domaine biomédical. En termes de cas d'usage, la SNOMED CT contenant plus de concepts morphologiques, topographiques et de néoplasmes que la CIM-11, elle offre la possibilité d'indexer et d'annoter un plus grand nombre de notions. Cependant, l'existence de concepts ambigus et peu informatifs peut être un frein à son interfaçage avec les utilisateurs. En revanche, le fait que la SNOMED CT décrive des concepts d'axes différents (y compris des modificateurs, tels que « aigu » ou « gauche ») rend cette ressource sémantique particulièrement utile à des fins d'interopérabilité (la SNOMED CT jouant potentiellement le rôle de pivot entre des ressources sémantiques représentant des concepts d'axes distincts). Notons enfin que les résultats obtenus dans ce travail méritent d'être approfondis par des analyses fouillées du contenu des ressources sémantiques pour apporter des conclusions plus complètes d'un point de vue qualitatif.

3.4.5 Cas d'usage alignement SNOMED CT / CIM-11 (LIMICS)

Le détail de l'étude est disponible en **annexe P4.5.**

3.4.5.1 Contexte et méthode

L'objectif de cette étude est de quantifier l'alignement de la CIM-11 avec d'autres terminologies disponibles sur le serveur HeTOP et d'analyser leurs couvertures respectives :

- évaluation de la couverture terminologique respective (capacité d'annotation) des principales terminologies de santé dans un entrepôt de données de santé (EDS) et dans une base de données bibliographique ;
- travail en fonction des effets recherchés (couverture de la SNOMED CT par rapport à toutes les autres, par exemple) ;
- précision de l'apport unique de chaque terminologie (concept unique annotant non retrouvé dans une autre terminologie))
- cas d'usage : annotation sur Entrepôt de Données de Santé (EDS) et base bibliographique.

3.4.5.2 Résultats

- **Alignements**

L'outil d'alignement associé à HeTOP a permis d'aligner 30 736 codes CIM-11 différents en exact match avec SNOMED CT, c'est-à-dire que les termes ou concepts alignés contiennent exactement les mêmes mots, ce qui permet de maximiser la précision, donc de minimiser les faux positifs, quitte à accepter plus de faux négatifs.

Tableau 12 : Alignements entre CIM-11 (59 000 termes) et les SOC de HeTOP

Terminologie	Nb concepts	Alignements manuels	Alignements automatiques	dont supervisés par un expert	% supervisé	Dont validés en « exact match »	% Exact match
MeSH	378 085	224	24752	12434	50,2	10747	43,4
SNOMED CT	326 946	207	30736	11000	35,8	9547	31,1
NCIT	140 319	82	13900	5659	40,7	4783	34,4
SNOMED int.3.5 VF	106 291	60	13087	5095	38,9	4701	35,9
MedDRA	68 140	7	13198	3793	28,7	3511	26,6
Orphanet	15 165	0	5849	2225	38,0	1884	32,2
CIM-10	19 858	0	7167	2069	28,9	1520	21,2
BNPC	93 890	8	2433	1943	79,9	1853	76,2
OMIM	16 709	16	2685	1215	45,3	1034	38,5
Médic./racines	91 058	18	1764	1645	93,3	1295	73,4
Disease Ontology	9 506	6	4546	1428	31,4	1277	28,1
FMA	104355	37	2951	479	16,2	396	13,4
LOINC	96374	3	2113	1151	54,5	1017	48,1
HPO	13725	4	2276	630	27,7	585	25,7
RadLex	46633	4	3009	612	20,3	567	18,8
ATC	6329	0	2398	2294	95,7	32	1,3

- **Couverture terminologique (Annotations)**

Les travaux ont été menés sur les traductions françaises des ressources sémantiques.

La SNOMED CT possède la plus grande capacité d'annotation des terminologies testées : 61 030 concepts trouvés. La CIM-11 arrive en 4ème position derrière le MeSH, MedDRA et le NCIT (14 704 concepts trouvés pour la CIM-11).

La CIM-11 possède environ trois fois plus de concepts que la CIM-10. Il est logique de trouver une capacité d'annotation nettement supérieure pour la CIM-11 par rapport à la CIM-10 : 14704 vs 5695. Ce facteur trois se retrouve également dans le nombre de concepts uniques annotant un document de l'entrepôt de données, sans aucun lien avec les autres terminologies : 7747 vs 2701. Ceci montre tout le progrès de cette nouvelle version par rapport à l'ancienne.

Chaque terminologie produit des singularités d'annotation montrant la complémentarité des différentes terminologies. Ces concepts uniques reflètent la spécificité des terminologies respectivement

Ces résultats démontrent l'action synergique des ressources sémantique sur l'annotation de documents de santé. Ceci renforce une vision multi-terminologique pour l'ensemble du système de santé français, et en particulier pour annoter de manière optimale les documents de santé d'un EDS.

Un travail qualitatif approfondi devrait être réalisé pour analyser les singularités d'annotation et déterminer celles qui sont opérationnelles et utiles. La comparaison SNOMED CT et CIM-11 devrait être également réalisée à domaine comparable (maladies, signes et symptômes, cause de mortalité)

En effet, SNOMED CT code des concepts qui sont très génériques et qui n'ont que peu de valeur opérationnelle à l'annotation (ex : « est un », « dossier », « par mois ») ce type de concept est également retrouvé dans la CIM-11 (ex : « droit », « gauche », « abdomen »). Ces concepts correspondent le plus souvent à des termes de post coordination.

Tableau 13 : Capacité d'annotation/indexation des différentes terminologies par rapport à l'entrepôt de données du CHU de Rouen (novembre 2019)

Terminologies	Total	Nb concepts trouvés	Nb concepts non trouvés	Nb concepts uniques sans liens avec les autres terminologies
Médic./racines	6165	4421 (71,71%)	1744 (28,29%)	1311 (21,27%)
MedDRA	68137	33783 (49,58%)	34354 (50,42%)	**20125** (29,54%)
FMA	17812	7883 (44,26%)	9929 (55,74%)	5557 (31,2%)
NCIT	80130	25720 (32,1%)	54410 (67,9%)	9090 (11,34%)
SNOMED CT	203282	61030 (30,02%)	142252 (69,98%)	**23826** (11,72%)
CIM-10	20053	5695 (28,4%)	14358 (71,6%)	2701 (13,47%)
CIM-11	55255	14704 (26,61%)	40551 (73,39%)	**7747 (14,02%)**
MeSH	185856	45428 (24,44%)	140428 (75,56%)	**15465** (8,32%)

3.4.5.3 Conclusion

La SNOMED CT ne s'aligne automatiquement qu'avec environ 50% des codes de la CIM-11, montrant que la CIM-11 ne peut être remplacée par la SNOMED CT.

Concernant la couverture terminologique de la CIM-11 au sein de 16,5 millions de documents de santé inclus dans l'entrepôt de données de santé, nous pouvons la mesurer à 21,54% (11904/55267).

La SNOMED CT possède la plus grande capacité d'annotation des terminologies testées, mais avec des concepts annotants qu'il faut approfondir pour déterminer leur caractère opérationnel.

Chaque terminologie possède des spécificités d'annotation non retrouvées dans d'autres terminologies, démontrant leur complémentarité.

La capacité d'annotation de la CIM-11 nettement supérieure à celle de la CIM-10, démontre tout le progrès entre ces deux versions.

La conclusion de cette étude de couverture terminologique renforce le choix stratégique du LIMICS d'avoir établi depuis plus de dix ans une vision multi-terminologique pour l'ensemble du système de santé français, et en particulier pour annoter les documents de santé d'un EDS.

3.4.6 Cas d'usage génomique (LIMICS)

Le détail de cette étude est présenté en **annexe P4.6.**

3.4.6.1 Contexte et méthode

Les psychoses chroniques, en particulier la schizophrénie, sont un enjeu majeur de santé publique. Elles sont fréquentes (2-3%), touchant chaque année en France environ 15 000 adolescents et jeunes adultes.

Dans ce cadre, le projet PsyCARE se propose de développer et de tester un ensemble d'outils innovants afin de faciliter l'accès aux soins, d'améliorer la détection précoce et d'offrir des programmes thérapeutiques personnalisés aux jeunes patients concernés, et ce, à l'échelle nationale. Le projet prévoit entre autres d'identifier des marqueurs génétiques de progression de la psychose et les cibles thérapeutiques correspondantes.

Dans le cadre de cette étude, les travaux du LIMICS visent à mesurer la capacité de plusieurs RTO à couvrir les différents aspects omiques, en insistant sur un cas d'usage particulier, celui de la psychiatrie dans le cadre du RHU (Réseau Hospitalo-Universitaire) PsyCARE.

Le LIMICS investigue par ailleurs, à travers cette étude, les performances de la SNOMED CT par rapport à plusieurs autres RTO en santé, dans ce contexte génomique.

Pour ce faire, l'équipe du LIMICS a réalisé :

- Une analyse du contenu des différentes ressources sémantiques ;
- Des travaux d'alignements entre les différentes RTO importantes dans le domaine (dont la SNOMED CT).

3.4.6.2 Résultats

a) Contenu des ressources sémantiques

10 ressources sémantiques ont été analysées :

NCBI Gene, NCBI protein, Uniprot, Human phenotype ontology (HPO), OMIM, Orphanet rare disease ontology (ORDO), Hugo gene nomenclature (HGNC), Gene ontology (GO), Medical Subject Headings (MeSH) et SNOMED CT.

Les 8 premières ressources ont la particularité de fournir des informations descriptives complémentaires sur un gène, sa localisation, son expression, sa traduction en protéine, ses possibles mutations, les possibles maladies qui lui sont associées, l'expression phénotypique de la maladie (lien génotype-phénotype), etc.

Ces ressources ont en commun d'être ouvertes et d'établir de nombreuses passerelles entre elles. Aucune de ces bases ne contient une information exhaustive, les passerelles (par l'identifiant normalisé du gène ou de ses synonymes, ou par le code de la séquence ADN, ARN ou de la protéine associée) permettent de lier les informations.

Le MeSH n'est pas une terminologie génomique. Il liste et indexe des gènes et des protéines sans les décrire. Notamment, il ne répertorie pas les mutations.

La SNOMED CT ne liste pas de gènes et pas de mutations. Elle liste les protéines en tant que substance à l'instar de MeSH. Il n'y a dans cette terminologie que quelques concepts de génomiques quand ils sont des signes de description de l'état de santé du patient dans un raisonnement clinique.
Le MeSH, comme la SNOMED CT, n'ont pas d'orientation génomique répondant aux besoins des utilisateurs généticiens. En effet dans une approche multimodale, plusieurs bases sont mobilisées pour chaque niveau d'exploration : génétique, épigénétique[40], transcriptomique, métabolomique, protéomique, et lipidomique[41].

[40] L'épigénétique est la discipline de la biologie qui étudie la nature des mécanismes modifiant de manière réversible, transmissible et adaptative l'expression des gènes sans en changer la séquence nucléotidique (Source Wikipédia).
[41] La lipidomique est la discipline étudiant les voie et réseaux de lipides cellulaires au sein de systèmes biologiques (Source Wikipédia).

D'autres passerelles vers des ressources dédiées peuvent être établies pour guider les cliniciens/chercheurs (interactions protéines/protéines, accès aux voies physiopathologiques impliquant le gène, classification des maladies liées au gène, etc.).

Le domaine de la génomique fait ressortir le besoin de données liées et d'interopérabilité pour les utilisateurs de ces données.

b) Alignements

Les alignements inter-RTO ont été réalisés sur le serveur de terminologie interlingue HeTOP (https://www.hetop.eu/hetop/fr/?q=&home) qui inclut 85 ressources termino-ontologiques.
La demi-matrice inférieure comprend les alignements en « exact-match » validés seulement par des membres du D2IM, (y compris ceux proposés par UMLS et les ressources termino-ontologiques de bio-informatique sélectionnées dans ce travail, qui en théorie ont été réalisés manuellement). La demi-matrice supérieure comprend tous les alignements en « exact-match » (c'est-à-dire en ajoutant les alignements automatiques qui ne sont pas encore validés).

Tableau 14 : Matrice d'alignement des ressources termino-ontologiques

Nombre de concepts alignés	Nbr concepts	HPO	OMIM	ORDO	HGNC	GO	Gène NCBI	Uniprot	MeSH	SNOMED CT
HPO	17083		821	1177	96	18	28	3	3043	7068
OMIM	25486	603		13590	17863	563	17677	16095	11408	6097
ORDO	13721	943	10905		4310	251	4191	4235	7686	8944
HGNC	41997	0	16224	4014		141	42125	20264	3258	1040
GO	45467	10	5	0	0		42	245	1971	1094
Gène NCBI	61563	0	16184	1	40007	0		19699	4030	471
Uniprot	20135	0	15081	3889	20048	0	19106		4454	1023
MeSH	277251	2015	5498	4545	18	479	0	93		55912
SNOMED CT	354042	5167	3842	6680	**9**	**334**	**0**	**45**	31280	

Si la SNOMED CT est correctement alignée avec les RTO des maladies rares (HPO, OMIM, Orphanet) ou les généralistes comme le MeSH, elle ne l'est pas du tout vis-à-vis des bases génomiques (de protéines et de gènes) (HGNC, GO, Gène NCBI et Uniprot). Ce résultat est important pour tous les spécialistes de bioinformatique ou plus généralement pour tous les chercheurs qui souhaitent travailler sur les données « clinomiques », à savoir les données mélangeant les données cliniques et omiques.

Ce résultat est à comparer aux alignements des RTO OMIM, Uniprot, HGNC et NCBI Gene entre elles, montrant la richesse des passerelles entre ces ressources.

Ces données « clinomiques » sont aujourd'hui à la base de toutes les études sur les données de vie réelle (« real world data ») issues des entrepôts de données de santé (EDS).

3.4.6.3 Conclusion

Cette étude a démontré qu'en fonction des cas d'usage dans le domaine des des sciences omiques, différentes bases peuvent couvrir les besoins des utilisateurs. Ces cas d'usage peuvent être de la description de gènes ou la représentation de liens entre des défauts génétiques ou protéiques et certaines conséquences phénotypiques. Concernant les bases et à titre d'exemples, pour l'étude des microARN (ARN n'ayant pas de produits protéiques contrairement à l'ARN messager), la base miRDB peut être interrogée. **L'analyse des gènes quant à elle ferait intervenir des ontologies telles que la base Gene NCBI, GO et HGNC**. Finalement, l'étude des maladies rares peut se faire à travers des RTO telles qu'Orphanet/ORDO ou encore OMIM). Ces RTO semblent centrales pour couvrir les cas d'usage du domaine omique.

Cette étude démontre par ailleurs la vitalité de la bioinformatique et la nécessaire interopérabilité entre les différentes bases afin de pouvoir représenter l'ensemble des concepts du domaine des sciences omiques.

Les alignements entre la SNOMED CT et les bases de protéines et de gènes (HGNC, GO, Gène NCBI et Uniprot) sont extrêmement faibles, tous inférieurs à 5%. Nous sommes dans des proportions bien inférieures à ce que nous constations quand nous nous intéressons aux terminologies cliniques : **nous pouvons dire que la SNOMED CT est en incapacité de couvrir le domaine de la génétique de manière générale** et que ce sont les bases/RTO étudiées dans cette étude (et peut-être d'autres dans un domaine très évolutif) qui en sont les référentiels. Par ailleurs et avec des bases telles que HGNC, GO, Gène NCBI et Uniprot, il semble possible de disposer des identifiants nécessaires à l'ensemble des utilisateurs.

La SNOMED CT au regard de sa représentation du domaine omique a plutôt tendance à se focaliser sur les aspects cliniques, comme en démontrent les résultats de cette étude.

La synthèse décrite ici est un instantané des modes de travail des chercheurs de PsyCARE. Nous pouvons raisonnablement penser que le mode de travail d'autres bio-informaticiens dans d'autres domaines n'est pas très différent, mais cela reste à investiguer.

3.4.7 Cas d'usage médicament Romedi (ISPED)

3.4.7.1 Contexte et méthode

Ce travail vise à étudier la couverture des connaissances liées aux médicaments par les ressources sémantiques de référence suivantes : la SNOMED CT, la CIM-11, l'ATC ainsi que les ressources disponibles au sein de la base de données publique des médicaments (BDPM) et le thesaurus des interactions médicamenteuses de l'ANSM.

Le choix s'est porté sur ces ressources pour des raisons propres à chacune :

- La SNOMED CT, étant la ressource sémantique la plus large et la plus utilisée dans le domaine biomédical et intégrant un axe « médicament » dans sa version récente ;
- La CIM-11 est la nouvelle version de la CIM proposée par l'Organisation Mondiale de la Santé (OMS) ;
- L'ATC est une classification spécifique du domaine ;
- La BDPM éditée par l'ANSM intègre la description des médicaments spécifiquement pour la France.

Deux approches distinctes ont été suivies :

1) Tout d'abord, les concepts de médicaments décrits dans l'ATC, la CIM-11 et la SNOMED CT ont été comparés d'après leurs termes normalisés en français et en anglais. Cette méthode vise à évaluer la complétude du domaine de connaissances du médicament par ces ressources sémantiques.
2) Ensuite, l'adéquation des termes décrits dans la SNOMED CT et dans la BDPM avec le langage utilisé par les professionnels dans les prescriptions en texte libre du CHU de Bordeaux a été étudiée. Pour l'évaluation de la BDPM, nous avons utilisé ROMEDI,[42] une ressource ouverte intégrant dans un graphe de connaissances les données de la BDPM, de l'ATC et du thesaurus des interactions de l'ANSM. Cette ressource, construite selon les principes du Linked Open Data, intègre également des liens vers des ressources externes (DrugBank, UMLS, wikipedia). Cette méthode apporte des éléments de discussion concernant les critères de précision et de véracité, en plus de la complétude qui est également considérée (cf. Annexe bibliographique P3.0).

3.4.7.2 Résultats

La première approche a permis de montrer que la couverture par la SNOMED CT et la CIM-11 des médicaments décrits dans l'ATC est globalement faible, n'atteignant pas 40% (tableau 1). Cela peut s'expliquer par l'existence dans l'ATC de concepts composites (ex : qui contiennent le mot « and » ou encore « incl. ») alors que la SNOMED CT et la CIM-11 les décrivent de manière atomique. De plus, certains concepts décrivant des catégories résiduelles (ex : contenant le terme « other ») ne se retrouvent pas dans la SNOMED CT et la CIM-11. Par ailleurs, nous avons observé que certains médicaments décrits dans l'ATC n'ayant pas été retrouvés dans la SNOMED CT y apparaissent néanmoins (ex : Histamine (phosphate) - 62666006 dans la SNOMED CT et Suramine (sodium) - 67928003). Aucun alignement n'a été obtenu pour ces concepts, car ils ne sont pas des descendants du concept Drug or Medicament (410942007) que nous avons utilisé pour identifier les médicaments dans cette ressource. Ces mêmes concepts sont décrits dans la CIM-11, mais n'ont pas non plus été « mappés » avec les concepts correspondants dans l'ATC à cause de la présence de parenthèses dans le libellé[43]. Ces « mappings » manquants pourraient être retrouvés en choisissant le concept plus général Substance (105590001) dans la SNOMED CT pour récupérer plus de concepts décrivant des médicaments et en appliquant une normalisation plus adaptée aux libellés de la CIM-11.

Par ailleurs, cette faible couverture des concepts de médicaments peut aussi être le résultat de niveaux de description différents dans les ressources sémantiques étudiées. Nous avons notamment rencontré des situations où un seul concept représenté dans une ressource sémantique était décrit par des concepts distincts dans une autre (ex : le concept SNOMED CT Profenamine (52850008) a pour synonyme Ethopropazine qui partage le même code

[42] https://www.data.gouv.fr/fr/reuses/romedi/
[43] Le phosphate d'histamine est un produit pour réaliser des témoins de tests cutanés. La Suramine est un médicament antiparasitaire utilisé dans la maladie du sommeil et aujourd'hui investigué dans l'autisme.

alors que la CIM-11 décrits deux concepts différents : Profenamine (938312196) et Ethopropazine (2089069547)), mais aussi reliés par ce lien de synonymie.

Notons enfin que chacune des trois ressources sémantiques considérées décrit des concepts qui lui sont propres en plus des concepts communs identifiés dans ce travail. Il est donc légitime de conclure qu'une meilleure couverture du domaine du médicament serait obtenue en utilisant l'ensemble de ces ressources qui apparaissent comme complémentaires.

Tableau 15 : Taux de couverture des concepts de médicaments entre l'ATC, la CIM-11 et la SNOMED CT (d'après les termes en anglais et en français)

Langue des termes	Couverture de l'ATC par la CIM-11	Couverture de l'ATC par la SNOMED CT	Couverture de la CIM-11 par la SNOMED CT
anglais	31,2%	39,0%	34,9%
français	27,3%	35,5%	31,2%

En ce qui concerne l'adéquation des termes, d'une manière générale, la base ROMEDI (BDPM combinée avec ATC) permet d'indexer plus de prescriptions (cf. tableau 2), retrouve plus de concepts et instancie plus d'occurrences de termes. Ce résultat était attendu puisque la BDPM décrit l'ensemble des médicaments prescrits en France suivant les dénominations spécifiquement françaises. Il est à noter que la BDPM indexe des termes très peu spécifiques tels que « autres » ou « sous cutané ». De même, on retrouve des termes génériques et non informatifs dans la SNOMED CT tels que « beta » ou « vaccin ». Enfin, la SNOMED CT indexe des termes plus génériques que la BDPM. A titre d'exemple, le terme « Tramadol » est indexé par la SNOMED CT alors que « Tramadol 50mg » est identifié par la BDPM. Cette différence est due à la spécialisation de la BDPM qui décrit les médicaments suivant leur dénomination, mais également avec les doses et les unités. En revanche, il faut noter que la SNOMED CT, bien que non spécialisée dans le médicament, intègre d'autres axes tels que les diagnostics et les maladies. Cette structuration suivant plusieurs axes permet d'envisager la description de liens utiles entre des entités disjointes (tels que les effets secondaires connus). Ainsi, il semble pertinent d'intégrer la BDPM et ATC avec la SNOMED CT afin de bénéficier de la structure de la SNOMED CT et de son caractère multiaxial.

Tableau 16 : Nombre de termes et de documents indexés suivant les différentes ressources et nombre de concepts distincts correspondants

Ressource sémantique	Nombre de termes	Nombre de concepts	Nombre de documents
SNOMED CT	524 088	1 091	171 538
ROMEDI (BDPM et ATC)	1 705 200	7 976	293 011

3.4.7.3 Conclusion

En conclusion, il apparaît que chacune des ressources analysées présente des atouts et des limites au regard des différents usages possibles. **Ainsi, il semble nécessaire de mettre à disposition l'ensemble de ces ressources et de les lier entre elles afin de couvrir le vocabulaire utilisé en routine (via la BDPM) et les liens vers d'autres types d'entités pouvant permettre d'envisager des usages tels que l'aide à la décision (interactions connues, effets secondaires possibles, contre-indications...). Cette approche pourrait favoriser une indexation large et efficace de l'information produite lors du soin et l'aide à l'intégration et l'interprétation automatique de ces informations pour l'aide à la décision autour de la prescription.**

Ce travail s'est focalisé sur le composant vocabulaire des ressources sémantiques et les deux cadres d'analyse ont permis de conclure que l'usage concomitant de la SNOMED CT, de la CIM-11 et de la BDPM est recommandé pour garantir une meilleure couverture du domaine biomédical.

En termes de cas d'usage, la BDPM associée à l'ATC permet de couvrir le champ des médicaments de façon plus large et spécifique que la SNOMED CT et offre la possibilité d'indexer et d'annoter un plus grand nombre de notions.

Cependant, le fait que la SNOMED CT décrive des concepts d'axes différents (y compris des modificateurs, tels que « aigu » ou « gauche ») rend cette ressource sémantique particulièrement utile à des fins d'interopérabilité (la SNOMED CT jouant potentiellement le rôle de pivot entre des ressources sémantiques représentant des concepts d'axes distincts et complémentaires). Ce caractère multiaxial pourrait également être exploité au sein d'applications d'aide à la décision dans le contexte de la prescription médicale. Cet usage nécessiterait un travail d'intégration de la BDPM avec la SNOMED CT afin de bénéficier du vocabulaire riche et précis disponible dans la BDPM et de la structuration de la SNOMED CT.

Notons enfin que les résultats obtenus dans ce travail méritent d'être approfondis par des analyses fouillées du contenu des ressources sémantiques pour apporter des conclusions plus complètes d'un point de vue qualitatif.

3.4.8 Cas d'usage médicament PsyHAMM (LIMICS)

Le détail de l'étude est disponible en **annexe P4.8.**

3.4.8.1 Contexte et méthode

Cette étude compare la SNOMED CT aux principales ressources onto-terminologiques traitant du médicament (ATC, NCIT, MeSH, CIM-11, Medicabase, etc.), ainsi que des ressources nationales sur le médicament, à savoir, en provenance de la base de données publiques du médicament (BDPM) :

- dénominations communes internationales (DCI) ;
- spécialités pharmaceutiques (codes CIS) ;
- présentations médicamenteuses (code CIP).

Ou de référentiels hospitaliers pour les unités communes de dispensation (UCD).

Cette comparaison est réalisée dans le cadre du projet PSYHAMM étudiant la prescription hors Autorisation de Mise sur le Marché (« hors AMM » par la suite) de médicaments psychiatriques. En effet ce projet requiert la modélisation du médicament français dans un contexte de prescription ainsi que des référentiels médicamenteux fiables.
Les objectifs de l'étude sont doubles :

- positionner la SNOMED CT dans les cas d'usage du domaine du médicament (avantages et limites) ;
- faire des recommandations sur un modèle de représentation du médicament utilisable en recherche.

Deux comparaisons ont été menées entre la termino-ontologie des médicaments développée dans le projet PSYHAMM et les principales terminologies généralistes (SNOMED CT, CIM-11, MeSH et NCIT) ou spécialistes (ATC).

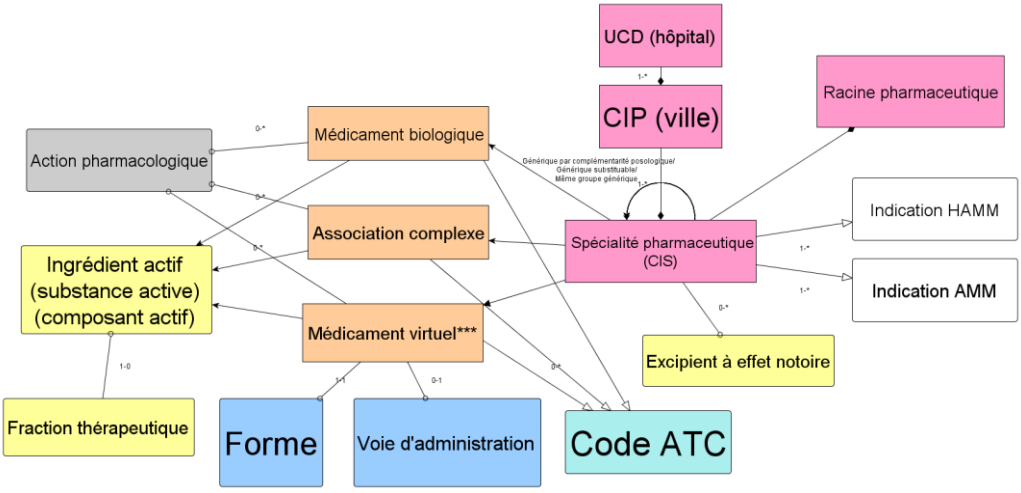
- la première a été menée pour mesurer les alignements terminologiques entre toutes ces terminologies ;
- la seconde a consisté à mesurer la couverture terminologique de ces mêmes terminologies au sein de l'entrepôt de données de santé du CHU de Rouen, qui contenait au moment de l'étude 16,5 millions de documents. Pour mesurer la couverture terminologique, nous avons utilisé l'annotateur sémantique ECMT.

3.4.8.2 Résultats

- **<u>Termino-ontologie du projet PSYHAMM</u>**

Une ressource sémantique a été bâtie sur la base du modèle médicament présenté ci-après.

Figure 12 : Modèle sémantique du médicament en France



Les données proviennent de sources publiques : base de données publique du médicament (ANSM), du référentiel du médicament virtuel (Médicabase) et du RCP.

La plupart des entités du modèle correspondent déjà à des ensembles de codes ou nomenclatures qu'il est nécessaire de mettre en correspondance. Les liens sémantiques sont définis explicitement ainsi que les cardinalités associées. Ces liens sont définis de la façon la "plus haute" possible afin de ne pas répéter l'information. Ce modèle est à la fois instancié sous forme de terminologie et d'ontologie pour des utilisations variées.

Les sources publiques souffrent d'un défaut de normalisation (formes, conditionnement notamment).

- **Alignement de la Terminologie PSYHAMM avec les terminologies de référence**

Le tableau ci-après présente les alignements entre les substances prescrites en France et les principales terminologies de santé :

- soit généralistes : CIM-10, CIM-11, MeSH, SNOMED CT et NCIT ;
- soit spécifique du médicament : ATC.

Tableau 17 : Alignements terminologiques sur les médicaments avec les terminologies de références (généralistes ou spécialisées) fondés sur les alignements HeTOP

Terminologies	Nb d'alignements exacts validés	Nb de composants alignés	% des alignts rapportés au nbre total de composants (3774)
MeSH (desc. + conc. Supp.)	1261	1253	33,20
MeSH (tout)	2912	1828	48,44
SNOMED CT	1870	1725	45,71
CIM-11	628	616	16,32
ATC	1328	1051	27,85
NCIT	1506	1494	39,59
Total (sans intersection)	1966	2027	52,10%

Sur les 3774 composants de médicaments disponibles en France issus de la base BDPM, seuls 1966 (52,1%) sont alignés (exact-match validé) vers CIM-11, ATC, SNOMED CT et MeSH).

Au total, en additionnant l'ensemble des alignements, on obtient un total de 2027 composants alignés sur 3774 (53,7%). Donc certains composants (46,3%) de médicaments ne s'alignent avec aucune Terminologie de référence. Ainsi, le MeSH dans son ensemble, incluant les MeSH concepts, et la SNOMED CT sont les deux terminologies généralistes qui s'alignent le mieux avec les DCI de la termino-ontologie de PSYHAMM (issue de la BDPM). Viennent ensuite le NCIT, puis l'ATC, pourtant une classification (de l'OMS (sic)) spécialisée sur les médicaments, et enfin la CIM-11.

Le modèle composé à partir des données publiques sur le médicament est donc supérieur à l'ensemble des terminologies testées en termes de complétude.

Ces résultats suggèrent le besoin d'une étude qualitative pour expliquer les manques des différentes terminologies de référence pour couvrir les différents niveaux de composants de la pharmacopée française.

- **Couverture terminologique (Annotation)**

La capacité d'annotation de la base PSYHAMM et des autres ressources ont été testées au sein des 16 millions de documents de santé présents dans l'entrepôt du CHU de Rouen

Le tableau ci-après présente les résultats obtenus en distinguant les terminologies du domaine médicament (DCI, nom commercial et ATC) des terminologies généralistes (dont la SNOMED CT et la CIM-11).

Tableau 18 : Couverture terminologique (capacité d'annotation) des terminologies de référence au sein d'un entrepôt de données

Terminologies	Total	Nb concepts trouvés	Nb concepts non trouvés	Nb concepts uniques sans lien avec les autres terminologies
Terminologies spécifiques du médicament				
DCI*	6165	3999 (64,87%)	2166 (35,13%)	**1081** (17,53%)
Nom commercial*	2580	1394 (54,03%)	1186 (45,97%)	**1393** (53,99%)
ATC	4007	2102 (52,46%)	1905 (47,54%)	204 (5,09%)
Total 3 bases	12 752	7 495 (58.78%)	5 257 (41.22%)	2 678 (21,00%)
Terminologies généralistes				
CIM-11	5125	1369 (26,71%)	3756 (73,29%)	111 (2,17%)
MeSH	21694	5352 (24,67%)	16342 (75,33%)	676 (3,12%)
SNOMEDCT	12969	2837 (21,88%)	10132 (78,12%)	369 (2,85%)
NCIT	12182	1908 (15,66%)	10274 (84,34%)	175 (1,44%)

** Issu de la BDPM*

De manière attendue, les base médicaments-spécifiques (DCI, noms commerciaux et classes ATC) sont les plus annotantes pour des documents médicaux : respectivement 64.87%, 54.03% et 52.46 % pour la base DCI, noms commerciaux et ATC. Au total en considérant ces trois bases disjointes (DCI, nom commerciaux et ATC) ensemble, on obtient une capacité d'annotation de 7 495 concepts (58.78% de l'ensemble des bases., avec 2678 concepts uniques.

Les singularités d'annotation (concepts annotants non-retrouvés dans d'autres terminologies) y sont également nombreuses par rapport à l'ensemble des bases testées : 1081 concepts uniques pour la base DCI et 1393 pour la base de noms commerciaux.

Ce nombre de concepts uniques est plus faible pour la base ATC (204). Une étude qualitative serait nécessaire pour étudier cette variation de capacité annotante par rapport à la base DCI qui est relativement proche. Ceci permettra notamment d'analyser les habitudes des professionnels de santé dans le report des traitements.

Les Terminologies généralistes n'atteignent pas cette capacité d'annotation et sont moins opérationnelles que les terminologies spécifiques.

La SNOMED CT n'a que 2837 concepts annotants sur 12969 (21.88%) concepts médicaments, avec seulement 369 singularités d'annotations. Ici aussi, une étude qualitative serait nécessaire pour préciser les limites de la SNOMED CT dans un cas d'usage médicament.

Le MeSH présente une plus grande capacité annotante que la SNOMED CT avec 5352 concepts annotants sur 21 694, et 676 singularités d'annotation.

3.4.8.3 Conclusion

La couverture sémantique issue de l'ontologie médicamenteuse composée dans le cadre du projet PSYHAMM à partir des données publiques de l'ANSM rassemblant tous les étages de description du médicament (de la Dénomination Commune Internationale (DCI) au nom de marque) est supérieure à toutes les terminologies de comparaison dans cette étude, qu'elles soient généralistes, comme la SNOMED CT ou le MeSH, ou spécialisées, comme l'ATC.

L'annotation par la SNOMED CT seule pour les médicaments n'est pas une solution opérationnelle. Il faut envisager une approche multi-terminologique, avec en premier lieu une liste à jour (y compris les Autorisation Temporaire d'Utilisation (ATU)) sur les noms commerciaux et les DCI, à partir des informations de la BDPM.

Il est important de noter que privilégier une approche multi-terminologique est nécessaire.

Afin de fiabiliser les analyses françaises sur le médicament, nous préconisons le renforcement de la BDPM pour qu'elle devienne la source d'une ontologie de référence du médicament librement accessible, sans préjuger de l'évolution des bases de données médicamenteuses déjà labellisées par la Haute Autorité de Santé, qu'elles soient privées ou publiques.

3.4.9 Besoins de codage dans les soins primaires (PML/ANS)

Les détails de l'étude sont présentés en **annexe P4.9**

3.4.9.1 Contexte et méthode

Le codage et la structuration des données de soins primaires sont des enjeux majeurs de fluidité et de continuité du parcours de soins. En effet la structuration des données permet de dématérialiser efficacement les échanges entre acteurs de soins primaires (médecins, infirmiers, pharmaciens et autres acteurs paramédicaux), mais aussi d'assurer une liaison entre dossier de ville et dossier hospitalier.
Des bases de données de soins primaires pourront être ainsi construites. Elles constituent un complément médical aux bases nationales telles que le SNDS pour apporter des d'informations cliniques.

Ces bases contiendront des données médicales :

- utiles en production de soins (sécurisation d'ordonnance, aide à la décision, programme de prise en charge coordonnée) ;
- utiles en exploitation de données (analyse de profils patients, sélection de cohortes, etc.) ;
- utiles en pilotage pour le professionnel de santé ou pour une structure de type maison de santé pluriprofessionnelle (pour produire des indicateurs d'activité relatifs au projet de santé de la structure).

Dans ce contexte, cette étude a été menée pour investiguer les besoins de codage et tracer les pistes d'amélioration de la structuration des données en soins primaires.

Elle poursuivait trois objectifs principaux :

1) Étayer les informations et contribuer à l'argumentaire autour de l'identification des domaines de connaissances d'intérêt et des ressources sémantiques attendues par les éditeurs de logiciels et les professionnels de santé ;
2) Définir les besoins de concepts à coder par les professionnels de santé en soins primaires ;
3) Comprendre et analyser les principaux freins et leviers à la structuration et au codage des données en soins primaires et identifier les besoins associés.

Les travaux ont été conduits en mobilisant des professionnels de santé des soins primaires et leurs représentants lors d'ateliers, d'entretiens semi-directifs et par l'envoi d'un questionnaire en ligne. Ils ont permis d'identifier plusieurs enseignements clefs qui permettent d'éclairer et de mieux interpréter les besoins exprimés.

3.4.9.2 Résultats

La structuration et le codage de leurs données constituent une préoccupation importante pour les professionnels des soins primaires. Ces deux notions ne sont néanmoins pas toujours comprises de la même façon puisque la structuration des données peut renvoyer à 3 sujets distincts :

1) la structuration des concepts utilisés en soins primaires (s'accorder entre professionnels de santé sur la façon de caractériser une information médicale ou sociale) ;
2) la structuration des logiciels métiers (où trouver une information) ;
3) la structuration de la donnée au sein du logiciel (le codage, ou association d'un code à un concept médical).

Au-delà de ce défaut de compréhension partagée, une certaine défiance vis-à-vis du codage persiste (perception négative issue notamment de la formation initiale hospitalière des professionnels, dans laquelle le codage est essentiellement associé à la facturation et à des technologies anciennes).

Le caractère chronophage ressenti du codage (accentué par l'inadaptation de certains outils aux soins primaires) s'ajoutant à ce défaut de compréhension et cette perception négative persistante explique le faible niveau de pratique.
Cependant, certains des professionnels de santé interrogés dans le cadre de ces travaux perçoivent la valeur des finalités de cette pratique : études épidémiologiques, travaux de recherche sur le domaine des soins primaires, optimisation de la transmission d'informations entre professionnels, aide à la prise de décision, …

La structuration et le codage de la donnée en soins primaires portent donc une ambition et un formidable potentiel puisque les généralistes réalisent à eux seuls 190 millions de consultations par an (Etude IQVIA 2018) contre 12,8 millions de patients hospitalisés dans l'ensemble des services en 2018 (source ATIH).

Les résultats de l'étude ont permis d'élaborer plusieurs préconisations :

A) Plusieurs actions auprès des professionnels de santé ou des éditeurs de logiciels doivent être envisagées de façon conjointe pour favoriser la structuration des données de santé et de leur codage dans le dossier du patient

Ces actions reposent sur trois leviers :

- Une prise en main de la gestion dématérialisée des données de santé (notamment de leur structuration et de leur codage) par les professionnels de santé ;
- Une simplification des activités de codage (notamment simplification des interfaces logicielles par les industriels éditeurs de logiciel) ;
- Une facilitation de l'accès des professionnels de santé à leurs données pour exploitation.

Ces actions nécessitent d'accompagner les changements nécessaires à la généralisation de l'activité de codage en soins primaires en mobilisant les professionnels de santé et leurs représentants, les éditeurs de logiciels et toute partie prenante de la chaîne de coordination de soins :

Tableau 19 : Objectifs, actions, acteurs et opportunités ANS

Objectifs	Exemples d'actions opérationnelles	Acteurs à mobiliser	Opportunité ANS
Contribuer à l'acculturation des professionnels de santé sur le sujet de la structuration des données			
Informer et former les professionnels	• « Evangéliser » la communauté Organiser des sessions orales et /ou présenter des posters de sensibilisation sur la structuration des données, les terminologies en usage, lors des temps d'échanges entre professionnels (congrès, journées professionnelles…) • A plus long terme, agir sur la formation Insérer dans la formation initiale et continue des professionnels (médecins et autres) une présentation de la structuration des données	• Représentants des professionnels de santé (URPS, sociétés savantes) • Organismes et institutions de formations universitaire et professionnels	• Construire des supports / modules de formation (MOOC, présentations, congrès, posters) liés au codage et à la structuration des données et les terminologies en usage (formation initiale et continue, congrès…) • Rédaction du guide d'utilisation des ressources sémantiques servant au codage et à la structuration. Diffusion de bonnes pratiques de codage.
Sensibiliser les soignants aux enjeux et aux finalités du codage	• **Capitaliser sur des retours d'expérience** Identifier et décrire des réalisations concrètes – « success stories » effectuées grâce à la structuration des données et à leurs échanges / exploitation (suivi d'une pathologie au sein d'une patientèle, rappel de prise en charge pour un patient donnée, aide à la décision grâce à l'intégration d'algorithmes, …) • **Conduire le changement** Identifier et prioriser les comportements à faire évoluer. Elaborer les messages au soutien du changement. Identifier les relais professionnels pour faire passer les messages au soutien de la généralisation du codage des données de santé.	• Représentants des professionnels de santé • Professionnels de santé • Editeurs de logiciel	• Elaboration de plans de communication Centraliser et diffusant et valorisant les « success stories » auprès des PS et des industriels • Construire un cercle d'expertise relai de savoir auprès de la communauté. • Produire une base de connaissance (guide sur le codage : quoi, pourquoi, comment)

Tableau 19 : Objectifs, actions, acteurs et opportunités ANS (suite)

Objectifs	Exemples d'actions opérationnelles	Acteurs à mobiliser	Opportunité ANS
Faciliter la prise en main de logiciels métiers par les professionnels			
Améliorer l'utilisation des logiciels pour le codage des données par les professionnels	• **Améliorer les interfaces utilisateurs** Impulser des travaux pour lister les besoins des professionnels de santé (champs à structurer, ergonomie, …) • **Inscrire les améliorations dans les feuilles de route des industriels.** Identifier des sponsors pour ces améliorations afin de les valider et de développer l'usage du codage	• Représentants des professionnels de santé • Professionnels de santé • Editeurs de logiciel	• Identifier les leviers d'intégration et d'améliorations dans les logiciels intégrer des exigences liées au codage dans les processus de labellisation des logiciels métiers • Promouvoir l'exploitation des données favorisée par les améliorations des interfaces logicielles

B) Réflexion sur l'adoption d'un vocabulaire complémentaire

En parallèle de ces actions, un travail d'instruction doit être initié pour identifier les terminologies couvrant les besoins rapportés par les professionnels des soins primaires. Les domaines sémantiques à investiguer par le CGTS sont ceux pour lesquels il a été rapporté un besoin de concept codé à intégrer dans les logiciels métiers utilisés par les professionnels rencontrés dans le cadre de cette étude. Les domaines identifiés sont présentés dans le tableau ci-après.

Tableau 20 : Synthèses des domaines

Thématiques ayant un besoin de codage non couvert actuellement	Typologie du besoin de codage exprimée
Examens de biologie pouvant être prescrits ou reçus par les praticiens	Résultats de biologie
Autres examens complémentaires pouvant être prescrits ou reçus par les praticiens	Prescription homogène pour les demandes d'examens complémentaires (imagerie, …) : type d'examen + organe concerné + latéralité
Prises en charges prescrites	Prescription homogène pour les soins à domicile ou autres interventions de professionnels auprès d'un patient (type d'intervention + type de professionnel + durée prescrite + fréquence
Prévention	Mention d'activités de prévention ou de programmes d'éducation thérapeutique
Médicament	Identifier des prescriptions par classes médicamenteuses
Vaccination	Mention de la valence vaccinale en prescription ou en suivi de vaccination
Données socioéconomiques	Eléments issus des registres sociaux, administratifs, liés à la situation familiale et professionnelle
Dispositifs médicaux	Prescription dans un format générique

La liste des domaines sémantiques conceptuels, bien que non exhaustive, est représentative du besoin des professionnels de santé des soins primaires.

3.4.9.3 Conclusion

Cette partie du chantier scientifique centrée sur les soins primaires, ne permet pas de répondre directement à la question de l'adoption d'une ou plusieurs terminologies de référence en soins primaires, notamment de la SNOMED CT.

Toutefois, elle apporte des arguments au débat, en montrant la grande distance actuelle des professionnels de soins primaires par rapport à une structuration des données de santé.

Cette étude permet de présenter une série d'opérations à mener successivement ou conjointement pour développer la structuration.

1. Acculturation des professionnels au codage/structuration des données.
2. Facilitation de l'action de codage et Développement des terminologies en usage (type CIM, CISP LOINC).
3. Facilitation de l'usage et exploitation des données structurées.

S'il paraît prématuré de parler adoption de terminologie de référence en soins primaires avant d'avoir développé la structuration des données de santé avec les terminologies déjà connues des professionnels de santé, cette étude a néanmoins permis d'identifier un certain nombre de domaines pour lesquels les professionnels de soins primaires ont exprimé un besoin de structuration. L'instruction de ce besoin pourrait démarrer sans délai et permettra d'étudier le positionnement de la SNOMED CT dans ces domaines par rapport à l'écosystème terminologique français de soins primaires.

3.4.10 Cas d'usage recherche (LIX/FX-Conseil)

Les détails de l'étude sont présentés en **annexe P4.10.**

La question de l'adoption de la SNOMED CT à des fins de recherche médicale est de plus en plus fréquemment posée.

Répondre à cette question suppose d'abord de rappeler à la fois les missions de l'ANS et le rôle des Terminologies en recherche et en informatique en général.

3.4.10.1 ANS et Recherche

L'activité de l'ANS et du CGTS autour des Terminologies de santé est légitimée par sa mission 1 « réguler » et liée à ses activités en interopérabilité en santé-social, que cette interopérabilité soit technique, sémantique ou syntaxique.

L'ANS n'a pas de mission de recherche, cette activité en santé étant notamment dévolue à des organismes tels que l'INSERM ou l'IRDES. De fait, savoir si oui ou non un référentiel est essentiel en recherche requiert des méthodes, des moyens et des critères d'évaluations dont certains sont applicables en interopérabilité et qui, comme mentionné plus haut dans le rapport, restent encore pour l'essentiel à formaliser.

Ll'ANS doit cependant se montrer exigeante vis à vis de la qualité des Terminologies qu'elle emploie, qui viennent pour l'essentiel du monde de la recherche ou qui doivent être avant tout évaluées en lien étroit avec elle.

3.4.10.2 Terminologies santé et activité de Recherche

Les précisions sur les liens entre ANS et le monde de la recherche étant faites, il est aujourd'hui évident que la recherche médicale ou sociale nécessite l'examen de données informatiques dont les volumes augmentent et pour lesquelles les besoins de production, de stockage, d'échange, de partage (et donc de codage et d'homogénéisation) sont de plus en plus importants.

A ce titre, toutes les Terminologies, qui sont des référentiels qui répondent parfaitement à ces 4 besoins informatiques fondamentaux, quel que soit leur niveau de formalisme (du glossaire à l'ontologie, cf. la biblio) sont utiles en recherche. **Réciproquement, la recherche médicale a tout intérêt à intégrer un maximum de Terminologies, plutôt que de se limiter arbitrairement à quelques-unes au risque d'ignorer des pans entiers de connaissance.**

Sur ce point, la recherche et l'interopérabilité ne se rejoignent pas complètement, l'une ayant besoin de maximiser les intégrations, l'autre ayant au contraire besoin de les rationaliser (notions de « *completeness* » et de « singularité » ou de « *conciseness* » vues plus haut) tant les coûts induits sont importants une fois les déploiements effectués.

Dans les deux cas cependant, les référentiels pris en compte doivent être corrects en adéquation avec les usages cibles (notion de « *correctness* » vue plus haut) et couvrir ensemble, une fois mis bout à bout, l'essentiel de la santé.

Pour donner un ordre d'idée, le CISMef de Rouen intègre 85 Terminologies santé et social différentes. L'UMLS aux Etats-Unis en gère 170. L'INSERM elle-même en produit ou en maintient plusieurs dizaines (MeSH, ORDO, Orphan Drugs, WHO-FIC...) pour ses seuls besoins, qui s'ajoutent à toutes celles qu'elle consomme.

La SNOMED CT est a priori utile, en combinaison avec d'autres Terminologies disponibles, en recherche médicale, sur des cas d'usage et/ou des projets qui devraient être identifiés.

Son adéquation aux usages cibles seule ou en combinaison doit être testée avant que l'Etat n'investisse de façon récurrente.

Par ailleurs, sur de tels projets de recherche, les investissements sont certainement rationalisables : par exemple par centralisation des traitements dans une base de données d'exploitation et donc limitation du nombre de licences à acquérir.

Il en va cependant de même pour beaucoup d'autres Terminologies, privées ou publiques, gratuites ou payantes, formelles ou non, dont les existences justifient qu'il ne soit pas fait un effort singulier sur l'une d'elles, ou sur la seule SNOMED CT.

Les limites des ressources incitent à diversifier l'offre : par exemple l'utilisation de la SNOMED CT en recherche est limitée par sa faible capacité de raisonnement, n'étant pas une ontologie formelle et par l'interdiction faites aux utilisateurs de pouvoir générer des œuvres dérivées avec de nouvelles relations ou des jeux de valeurs mixtes (cf. l'étude bibliographique et l'étude juridique).

3.4.10.3 Ontologies et découvertes issues de la recherche

Les ontologies permettent de représenter des concepts d'un champ de connaissance et les liens sémantiques qui les relient entre eux. En ce sens, ce sont des Terminologies dont le formalisme est le plus abouti.

Elles marquent un état de connaissance donné à un moment donné sur un sujet donné. Elles ne sont donc pas, par définition, porteuses des éléments de connaissance restant à découvrir.

En somme : ce qui est déjà formalisé n'est plus à découvrir. Aucune Terminologie, y compris la SNOMED CT, n'échappe à cette remarque.

Dès lors, le but de la recherche est d'étendre le champ de connaissance d'un domaine donné et, in fine, consciemment ou non, de compléter les ontologies qui s'y rapportent. La recherche ne doit donc pas, par principe, se contenter, dans leurs concepts, mais aussi dans leurs formes (organisations générales, niveaux de formalisme DL), des ontologies existantes.

Disposer d'ontologies permet à la recherche de ne pas redécouvrir « les bases ». Pour aller plus loin, la recherche dispose aujourd'hui des techniques informatiques (machine learning/deep learning for knowledge extraction, « ontology learning ») dont elle peut se servir pour découvrir de nouvelles corrélations. Ces découvertes, une fois qualifiées formellement, peuvent servir à enrichir les ontologies incomplètes qui ont servi de bases de départ, dans un cercle vertueux correspondant à l'amélioration continue de la connaissance en général.

3.4.10.4 Focus sur les techniques d'intelligence artificielle d'extraction de la connaissance (word embeddings ou prolongement lexical)

Dans le cadre de cette étude, l'équipe du LIX, s'est focalisée sur une technique d'extraction de la connaissance en essor (*word embeddings*) dans le domaine médical pour, évaluer les potentialités de cette méthode afin de la positionner comme technique d'extraction de la connaissance à partir de corpus documentaires au côté des techniques recourant à des ressources sémantiques.

Les techniques d'extraction de la connaissance de type *word embeddings*[44] (ou prolongement lexical) ont commencé à montrer leur plein potentiel en 2013 avec des applications de traitement automatique du langage naturel (NLP).

Le développement rapide des méthodes et des outils d'intelligence artificielle (IA) au cours des dernières années, associée à la recherche émergente en auto-encodeur[45] base du « prolongement lexical » ont intensifié le rythme de développement des applications à valeur ajoutée dans différents domaines, dont celui de l'extraction de connaissances médicales. Les percées successives de techniques de « prolongement lexical » réalisées par les laboratoires de recherche ont progressivement abouti à des modèles de connaissances opérationnels offrant des avantages certains à leurs utilisateurs finaux.

Dans ce domaine, des avancées scientifiques majeures sont observées depuis 2019 transformant les tâches de NLP et ouvrant de nouvelles voies dans l'exploitation et le traitement de données médicales y compris l'enrichissement de Terminologies médicales.

[44] Principe du prolongement lexical : déterminer une représentation des termes par un vecteur numérique de dimension K (paramétrable), en tenant compte de son contexte (fenêtre de voisinage V dont la taille est paramétrable).
[45] Les auto-encodeurs sont des algorithmes d'apprentissage non supervisé à base de réseaux de neurones artificiels, qui permettent de construire une nouvelle représentation d'un jeu de données. Généralement, celle-ci est plus compacte, et présente moins de descripteurs, ce qui permet de réduire la dimensionnalité du jeu de données. L'architecture d'un auto-encodeur est constitué de deux parties : l'encodeur et le décodeur.

L'extraction de connaissances à partir d'ensembles de données textuelles non structurées, comme les documents de santé et les rapports médicaux, peut être d'une importance cruciale pour de nombreuses applications dans le domaine de la santé. Les établissements de soins de santé continuent de faire face à un défi majeur pour extraire et enrichir les connaissances contenues dans des ensembles de données médicales non structurées et structurées stockées dans des référentiels numériques en constante expansion des dossiers de santé électroniques (DSE).

Des progrès ont été réalisés dans la modélisation de textes médicaux non structurés en représentations signifiantes, qui pourraient être utilisées pour identifier des liens cachés entre des concepts (générateurs d'inférences) ou pour développer des systèmes d'aide à la décision entièrement automatisés.

Le « prolongement lexical est la méthode la plus communément acceptée pour créer ces représentations. Historiquement, cette méthode prend ses racines à la fin des années 1960 avec le développement de modèle d'espace vectoriel multidimensionnel pour la représentation de l'information. Plus tard, Bengio et al. [2003] ont proposé de rationaliser par réduction, le nombre de dimensions de l'espace vectoriel de représentation des mots dans leurs contextes en introduisant une représentation distribuée des mots.

Un modèle de « prolongement lexical » est donc essentiellement une représentation de données dans laquelle des mots ou des documents individuels sont représentés comme des vecteurs numériques à valeur réelle dans un espace vectoriel prédéfini, où les mots ayant une signification similaire ont une représentation similaire, se rapprochant ainsi les uns des autres dans l'espace vectoriel. Ils peuvent être construits sur des corporal textuels brut non structuré ou des données lexicales structurées.

L'interaction de ces techniques avec des ressources termino-ontologiques (RTO) via « le prolongement lexical » est une approche prometteuse pour exploiter des ensembles de données textuelles médicales structurées ou non dans le domaine de la recherche médicale. L'exploration des connaissances à partir de ces ensembles de données à l'aide de « prolongement lexical », d'une manière approfondie, pourrait alimenter en retour le développement de ces terminologies et ontologies, créant ainsi des œuvres dérivées.

Au total, ceci rendrait les processus de raisonnement clinique, diagnostique et la génération d'inférence médicale plus rapides et plus efficaces.

Pour résumer, les chercheurs entraînent des « workflow » de « prolongement lexical » avec des Terminologies en entrée ou extraient des terminologies ad hoc via ces techniques.

Grâce à une revue de bibliographie (Détail en **annexe p 4.10**), nous arrivons à la conclusion que la structuration et l'organisation des connaissances à partir d'un ensemble de données (données médicales structurées ou non) pourraient être réalisées dans un schéma complémentaire entre annotation terminologique couplées avec des techniques de « prolongement lexical » judicieusement sélectionnées.

Elle démontre la possibilité de construire des terminologies spécifiques à la recherche, représentant une alternative à l'utilisation de terminologies telle que la SNOMED CT et permettant d'enrichir des ontologies existantes en créant des œuvres dérivées issues d'un processus de génération d'inférences.

Ce travail pointe également que les techniques de traitement de l'information évoluent très vite. Il est donc nécessaire de mettre en place une veille pour anticiper les révolutions technologiques qui ne manqueront pas de remettre en cause les choix sémantiques. Par exemple l'enrichissement de terminologies ou la création d'ontologies à partir de ressources existantes requièrent de pouvoir créer librement des œuvres dérivées.

3.4.11 Résumé : ce que les différents cas d'usage étudiés nous apprennent sur la SNOMED CT

Tableau 21 : Chantier scientifique

Avantages	Inconvénients
- La SNOMED CT est la terminologie présentant le plus grand nombre de concepts dans un large panel de cas d'usage. - La SNOMED CT apparaît comme utile au regard de sa combinaison possible avec d'autres terminologies - La SNOMED CT présente une grande capacité d'annotation.	- La SNOMED CT est moins performante en termes de couverture et d'annotation que les terminologies de domaines (microbiologie, anatomie, SLA, médicament, etc.). Bien que permettant de couvrir les aspects classiques du domaine médical, elle ne peut couvrir les concepts spécifiques de chaque domaine. - La SNOMED CT annote des concepts non-opérationnels (de type « est un »). - La SNOMED CT ne pourra pas remplacer la CIM-11. Elle serait complémentaire de la CIM-11 sur le cas d'usage interopérabilité.

4 DISCUSSION - SYNTHESE DES RESULTATS

4.1 Positionnement relatif de la SNOMED CT en fonction du cas d'usage (chantier scientifique)

Quatre grandes familles de cas d'usage peuvent être envisagées pour les Terminologies (Cf. chantier bibliographique).

Les travaux scientifiques et bibliographiques ont permis de positionner la SNOMED CT dans l'ensemble de ces cas d'usage. Le tableau 22 (Cf. Infra) résume les différentes conclusions dans cet axe.

4.1.1 Interface utilisateur

La SNOMED CT se positionne le moins bien dans le cas d'usage interface de structuration de la donnée, car elle fait face soit à des terminologies de domaine plus complètes (NCBI Taxonomy, FMA, OntoParon), soit à des terminologies en usage (CIM, CCAM, NABM, LOINC, ADICAP, etc.).

Dans l'échantillonnage de cas d'usage analysés au cours de cette étude, elle se positionne comme inférieure aux terminologies de référence en termes de couverture (microbiologie, anatomie, maladie de Charcot, médicament ou génomique). Elle est équivalente à la CIM-11 dans le domaine de la cancérologie.

En soins primaires, il est plus nécessaire de développer un accès facilité aux Terminologies en usage (CIM/CISP/ATC/Vaccins) voire une acculturation au codage des données. Il n'y a pas de besoins immédiats concernant la SNOMED CT.

4.1.2 Interopérabilité

La SNOMED CT se positionne en interopérabilité en tant que terminologie pivot. Elle serait souvent présente en deuxième couche en alignement avec les Terminologies d'interface et en complémentarité avec d'autres ressources.

A noter que les différences de lignes éditoriales entre les différentes terminologies font que le travail d'alignement peut être conséquent à réaliser et difficile à maintenir.

Dans ce cas d'usage, il est plus simple et moins couteux d'utiliser les terminologies d'interface en interopérabilité quand cela est possible.

4.1.3 Annotation, indexation, recherche

C'est dans ce cas d'usage que la SNOMED CT se positionne le mieux au cours de cette étude. Ce cas d'usage met en valeur sa richesse. Néanmoins ici aussi, elle reste complémentaire d'autres ressources pour une annotation optimale, chaque ressource sémantique possédant ses singularités d'annotation comme l'ont démontré les multiples comparaisons d'annotation de la SNOMED CT aux cours des études menées dans le cadre de ce rapport.

Dans ce cas d'usage, sont avantagées les terminologies présentant un large volume de concepts avec une grande variété de synonymie et une arborescence optimisée.

Tous les contributeurs de l'étude font ressortir l'intérêt d'une approche multi-terminologique pour une extraction de connaissance optimale à partir d'entrepôts de données.

4.1.4 Raisonnement, déduction inférences

La SNOMED CT n'a pu être étudiée spécifiquement dans ce cas d'usage au cours de cette étude.

Le chantier bibliographique rappelle ses limites à ce niveau, la SNOMED CT n'étant pas une ontologie complètement formelle.

Par contre l'étude réalisée par les chercheurs de l'école Polytechnique la positionne comme corpus de base au côté d'autres terminologies pour servir à extraire de la connaissance de corpus textuels et générer de nouvelles connaissances.

A ce titre, la SNOMED CT pourrait être enrichie ad hoc avec restructuration de ses liens sémantiques pour générer des ontologies de domaine plus adaptées aux cas d'usage cibles.

Mais, ici c'est le modèle propriétaire de la SNOMED CT qui est un facteur limitant à son emploi dans des workflow d'IA de type prolongement lexical.

Le tableau ci-après synthétise le positionnement de la SNOMED CT au cours de cette étude.

Tableau 22 : Chantier scientifique - Positionnement relatif de la SNOMED CT

Cas d'usage	Positionnement relatif de la SNOMED CT	Etude d'appui
Interface utilisateur Encodage Codage manuel	Couverture inferieure aux Terminologies de domaine Couverture inferieure de SNOMED CT et CIM-11 vs CIM-O3 Pas de besoin	*Microbiologie* *Anatomie* *Maladie de Charcot (SLA)* *Génomique* *Oncologie* *Soins primaires*
Interopérabilité Alignement de ressources sémantiques	Complémentaire de ressources en usage Pas de besoin	*Alignement CIM-11 / SNOMED CT* *Médicaments* *Oncologie* *Génomique*
Annotation, Indexation, Recherche	Grande capacité d'annotation Complémentaire d'autres ressources Pas de besoin	*Maladie de Charcot (SLA)* *Médicament (ISPED)* *Oncologie* *Génomique*
Raisonnement, déductions et inférences Système d'aide à la décision	Limitée si utilisée seule Utile en complément d'autres techniques	*Bibliographie* *Recherche (Word Embeddings)*

4.2 Avantages et inconvénients de la SNOMED CT

Le tableau ci-après rassemble les observations réalisées au cours des différents chantiers.

Tableau 23 : Synthèse des avantages/inconvénients de la SNOMED CT

Avantages	Inconvénients
Retour d'expérience international	
Ecosystème SNOMED CT performant	Investissement financier conséquent Pas de remplacement des terminologies existantes (CIM-10, actes, biologie, médicament…) Effort conséquent d'intégration et d'accompagnement
Chantier juridique	
Modèles de licences clairement définies Existence d'un outil dédié à la gestion du parc de licences (MLDS)	Modèle de propriété intellectuelle fermé faisant figure d'exception dans le domaine des vocabulaires d'interopérabilité Ambiguïté de certaines clauses des licences Clause de résiliation très contraignante (système de location de données) à négocier pour ne pas impacter le tissu industriel français en cas de retrait
Chantiers bibliographique et scientifique	
Terminologie multi-domaines présentant un grand nombre de concepts Terminologie pivot sur laquelle s'articulent des terminologies plus complètes Grande capacité d'annotation Terminologie complémentaire	Couverture partielle, inférieure à des terminologies de domaine La SNOMED CT n'est pas une ontologie complètement formelle, son utilisation en raisonnement est limitée Terminologie internationale présentant des concepts non opérationnels sur le plan local Limitation de son usage en recherche par l'absence de possibilité de créer des jeux de valeurs mixtes multi-terminologies par les utilisateurs ou de recréer des liens sémantiques

4.3 Recommandations

Au vu des résultats de cette étude, il est prématuré d'adopter la SNOMED CT au niveau national pour un déploiement à court terme.

Afin de développer l'interopérabilité des données de santé et leurs usages en recherche, 4 recommandations peuvent être faites :

1. **Constituer un corpus sémantique national cohérent à partir des Terminologies en usage en France :**
   - *L'écosystème sémantique des Terminologies en usage en France est déjà riche, c'est son hétérogénéité qui nuit à l'interopérabilité : Il faut l'organiser et le structurer. Il faut se concentrer sur la mise en qualité de ce corpus sémantique[46] existant ainsi que son usage sur le terrain. En effet très peu des terminologies du catalogue du CGTS sont disponibles sous format normalisé. Certaines ont besoin d'une mise à niveau par leurs UP (LPP (CNAM) pour coder les dispositifs médicaux). De manière générale, les terminologies de la CNAM doivent intégrer le catalogue du CGTS afin que leur format de distribution soit standardisé ;*
   - *Cette offre doit être organisée et animée par l'État, en particulier l'ANSM, les directions centrales et la CNAM qui doivent garantir et développer les mises en œuvre de leurs référentiels sémantiques essentiels dans les échanges de santé (Référentiels DM, actes, médicaments, référentiels médicaux, référentiels du handicap) dans des cas d'usage prioritaires (volet numérique du Ségur de la santé, e-prescription, ENS, DMP, DCC, médico-social) ;*
   - *Un effort immédiat doit être réalisé sur les Terminologies en usage et sur les nouvelles ressources sémantiques de l'OMS (CIM-11, ICF et ICHI) dont les usages en interopérabilité vont bien au-delà de leur usage classique en statistique. Elles sont à considérer en priorité. La place de la SNOMED CT fera jour après ce travail indispensable ;*
   - *En amont, la constitution de ce corpus dépend de la fiabilité des unités de production sémantiques chargées de constituer, de traduire, d'aligner et de maintenir les ressources. A ce titre, les unités de production institutionnelles (ANSM, CNAM, ATIH, etc.) doivent être exemplaires en termes de qualité. Les unités de production doivent s'engager dans un processus de long terme afin de pérenniser les ressources dont elles sont responsables. L'état doit aussi s'engager dans un accompagnement des Unités de production pour les aider dans leurs tâches et la constitution d'un modèle de production pérenne. Ceci est particulièrement crucial pour les ressources essentielles comme celles de l'OMS ;*
   - *En aval, la constitution de ce corpus doit être accompagnée d'opérations de communication et de formation pour fédérer tous les acteurs de l'interopérabilité autour de cet espace de confiance de publication des référentiels sémantiques.*
2. **Pour toute nouvelle ressource sémantique, il est nécessaire d'évaluer avant de préconiser :**
   - *L'étude montre les forces et faiblesses techniques de la SNOMED CT ainsi que les problématiques juridiques à résoudre ;*
   - *Il en va de même avec toute nouvelle ressource candidate à être terminologie de référence ;*
   - *Le travail sur le volet numérique du Ségur de la santé permettra de poursuivre l'étude du positionnement de la SNOMED CT sur des cas d'usage prioritaires.*

[46] On rappellera que les deux terminologies les plus utilisées en France CIM-10 à usage PMSI et NGAP sont distribuées en PDF.

3. **En cas d'adoption de la SNOMED CT, des ressources sont nécessaires (NRC/CGTS) pour la maintenir dans un écosystème terminologique fourni, en supplément aux ressources nécessaires pour la maintenance du corpus de terminologies en usage :**
   - *L'étude montre que la SNOMED CT ne remplacera pas toutes les Terminologies et l'effort associé à son déploiement s'ajoutera aux travaux du CGTS sur les autres ressources primordiales (Volet numérique du Ségur de la Santé, ENS, e-prescription, publication des ressources de l'OMS, parcours oncologie, etc.) ;*
   - *Une adoption de la SNOMED CT implique de négocier d'une part les clauses ambiguës et celles liées à la résiliation et d'autre part, un positionnement adapté en fonction de ses forces et faiblesses ;*
   - *L'adoption de la SNOMED CT implique une participation active à la gouvernance de SNOMED International ;*
   - *A priori une estimation de 2 ETP dédiés a été réalisée pour assurer le déploiement et l'animation de la communauté SNOMED CT en France.*
4. **La technique évolue très vite : un système de veille est nécessaire pour anticiper les révolutions techniques qui ne manqueront pas de remettre en cause les choix sémantiques :**
   - *Il est nécessaire de participer aux gouvernances des instances normatives internationales pour porter les intérêts français, pérenniser des formats d'échanges pertinents en France et identifier des partenaires sur les cas d'usage prioritaires ;*
   - *De même, il est nécessaire de participer aux gouvernances européennes qui définissent les standards et profils d'échange transfrontaliers en Europe ;*
   - *Ceci implique donc la participation active aux instances de pilotage des référentiels retenus : LOINC, CISP, OMS, EMDN/CND.*

4.4 Scenarios pour le futur

Afin d'éclairer le lecteur, l'ensemble des décisions d'adoption et leur impact terrain a été mis en scénarios :

- **Scenario 1** : Adoption de la SNOMED CT
- **Scenario 2 :** Pas d'adoption de la SNOMED CT
- **Scenario 3 :** Poursuite de l'évaluation de la SNOMED CT

A noter que dans ces trois scénarios :

- L'offre actuelle de Terminologies sera renforcée. Le CGTS en assurera la publication ;
- La France devra déployer les terminologies de l'OMS ;
- Les terminologies réglementaires et/ou impliquées dans la chaîne de facturation (ex : CIM, CCAM, NGAP, NABM, LPP, CIP/UCD pour le médicament) sont incontournables.

4.4.1 Scenario 1 : Adoption de la SNOMED CT

Dans le cas où la DNS se positionne en faveur d'une adoption nationale d'une terminologie pivot et de la SNOMED CT en particulier, le processus d'achat doit être déterminé sur les bases d'un cahier des charges.

Ce cahier des charges exprimera le besoin de l'écosystème de l'interopérabilité en termes de ressources sémantiques, d'exploitabilité, d'accessibilité et de souveraineté des données.

Dans cette phase, il devra être vérifié :

- L'adéquation de la licence par rapport à un cahier des charges français respectant notamment la souveraineté des données de santé ;
- Le respect des exigences techniques et de qualité.

En adoptant la SNOMED CT, la France deviendrait ainsi un membre de SNOMED International. L'Etat désignera un NRC selon les critères définis par SNOMED International (le NRC peut être porté soit par une agence gouvernementale soit par une société approuvée par le gouvernement et dont les responsabilités et les activités sont pertinentes) :

- L'Etat contractualisera avec SNOMED International conformément au cahier des charges :
  - Les clauses contractuelles posant question ayant été clarifiées et négociées (notamment les clauses 2 et 3 sur les droits de modification et de publication et clause 5.6 concernant les conditions de résiliation) ;
  - Les termes de la licence nationale ainsi que les droits de propriété intellectuelle des traductions et des extensions nationales devraient également être négociés (clauses 3.5 et 3.6).

Une fois les problématiques juridiques résolues, la phase opérationnelle pourra débuter.

Dans la perspective du cadre d'interopérabilité :

- Le NRC sera mis en Run avec les ressources nécessaires et suffisantes pour assurer le déploiement et la mise en usage de la SNOMED CT. Le NRC sera l'unité de production de la SNOMED version française ;
- La SNOMED CT sera traduite en capitalisant sur les efforts déjà faits (CHU Rouen, traduction ANS/CGTS, effort de la francophonie pour traduire la SNOMED CT). Une unité de production SNOMED VF sera constituée par appel d'offre : traduction, synthèse et remontées des besoins, formations, etc. ;
- Le **cadre d'interopérabilité** intégrera la SNOMED CT. Son positionnement par rapport aux Terminologies en usage sera à déterminer (remplacement ou superposition). Ce travail sera clé pour déterminer le volume de travail d'alignement à produire ;
- Un plan de déploiement doit être élaboré selon les préconisations d'ASSESS CT. Il devra trouver son financement. Les cas d'usage seront priorisés pour respecter un développement progressif ;
- Les éditeurs de logiciels pourront intégrer gratuitement la SNOMED CT dans leurs logiciels ;
- Une **formation et un accompagnement** des professionnels pour intégrer la SNOMED CT dans leurs logiciels sera à assurer ;
- Des **extensions nationales** seront à créer et gérer par le NRC local ;
- A terme, la SNOMED CT sera le choix préférentiel pour toute instruction de nouvelles terminologies, les autres ressources sémantiques seront examinées en cas de non-couverture du besoin par cette ressource.

4.4.2 Scenario 2 : Pas d'adoption de la SNOMED CT

Dans ce cas, la France n'est pas adhérente à SNOMED International. Les industriels désireux d'intégrer la SNOMED CT restent libres de le faire et doivent payer des licences affiliées selon leurs besoins.

Dans la perspective du cadre d'interopérabilité :

- Celui-ci s'appuiera en priorité sur les terminologies en usage ;
- Le budget dédié au CGTS servira à renforcer les terminologies en usage et à constituer un corpus national cohérent de Terminologies (instruction de nouvelles ressources) ;
- L'instruction de nouvelles ressources se fera à l'aide d'évaluations. La SNOMED CT continuera à être évaluée dans le cadre de l'instruction de nouveaux cas d'usage ;
- Les normes d'interopérabilité qui sont fournies avec des pointeurs SNOMED CT (ex : FHIR) devront être profilées pour pointer également sur d'autres ressources sémantiques. Le portage de terminologies multiples dans les normes d'échanges aura l'avantage de porter une interopérabilité nationale et internationale. Les alignements seront réalisés pour les cas d'usage internationaux. Il faudra tenir compte des standards européens (par exemple IDMP pour les médicaments et CND/EMDN pour les dispositifs médicaux).

4.4.3 Scenario 3 : Poursuite de l'évaluation de la SNOMED dans le cadre de projets de recherche et d'expérimentations

Entre une non-adoption et une adoption nationale d'emblée, toute une gamme de scénarios peut être envisagée :

- Créer un environnement favorable à l'utilisation de la SNOMED CT sur le sol français afin de pouvoir analyser des retours terrain : produire une traduction qualifiée et la mettre à disposition de tout industriel désireux d'utiliser la SNOMED CT. Ceci implique une négociation avec SNOMED International pour diffuser une version française sans adhérer à SNOMED International et une stimulation des industriels à produire des retours d'expérience ;
- Organiser une expérimentation de terrain sur un projet où la SNOMED CT peut faire face à des alternatives ;
- Adopter la SNOMED CT et faire participer les industriels utilisateurs de la SNOMED CT aux frais de souscription et de mise en place du centre collaborateur. Ceci implique également de négocier une licence acceptable du point de vue propriété intellectuelle pour ne pas pénaliser les industriels et utilisateurs désireux de se retirer de la SNOMED CT. Une adoption nationale sera envisagée si la SNOMED CT est réellement en usage en France et que les conditions juridiques de son utilisation sont réunies ;
- Suivre l'expérimentation allemande.

Dans tous les cas évoqués, la problématique de la souveraineté des ressources sémantiques (accord sur les droits de propriété intellectuelle concédés par SNOMED international) utilisées en France doit être résolue dans la perspective d'une adoption nationale éventuelle de la SNOMED CT.

Dans la perspective du cadre d'interopérabilité, nous sommes très similaires au scénario de non-adoption :

- Focalisation sur les terminologies en usage ;
- Priorité à la mise en qualité du catalogue ;
- La SNOMED CT est instruite comme candidate au même titre que d'autres ressources terminologiques ;
- Choix des normes d'interopérabilité les plus ouvertes embarquant plusieurs systèmes de codage pour maximiser l'interopérabilité internationale.

Un suivi des expérimentations[47] et/ou des usages de la SNOMED CT sera effectué pour réévaluer le bénéfice d'une adoption nationale.

[47] http://www.journee.snomed.fr/JFSCT2019-01-Sylvia_THUN.html minute 2:20 : L'Allemagne a décidé en 11/2019 de lancer une procédure d'évaluation de la SNOMED CT avec une première phase de 1-2 ans impliquant d'abord le Ministère de la Science et de l'Education sur des projets de recherche, avant de prendre une décision d'adoption.

5 LISTE DES ANNEXES

Tableau 24 : Liste des annexes

Axe	Nom	Description
Retour d'expérience international	P1.0 : Déploiement et cas d'usage de la SNOMED CT : retours d'expérience de centres de gestion Terminologiques européens.	Cette étude présente les retours d'expérience d'utilisation de la SNOMED CT de 4 pays : les Pays-Bas, l'Allemagne, la Norvège, la Pologne (Etude PML/ANS).
Juridique	P2.0 : Positionnement juridique de la SNOMED CT dans l'écosystème sémantique français : propriété intellectuelle et sécurité juridique.	Cette étude positionne la SNOMED CT dans l'écosystème des Terminologies en termes de propriété intellectuelle et de sécurité juridique (Etude DJUR/ANS, KGA, PML).
Bibliographie	P3.0 : Revue de la littérature sur les méthodes d'évaluation des ressources sémantiques et le statut de l'évaluation de la SNOMED CT.	Cette étude, présente un état des lieux de la littérature sur les méthodologies d'évaluation des ressources sémantiques ainsi qu'un statut sur l'évaluation de la SNOMED CT (Etude ISPED, LIMICS, IQVIA).
Scientifique	P4.1 : Comparaison SNOMED CT et NCBI Taxonomy par rapport à un référentiel en usage à l'AP-HP.	Cette étude, évalue les performances de la SNOMED CT sur le cas d'usage microbiologie par rapport au besoin de terminologie de référence exprimé par l'AP-HP (Etude AP-HP, PML).
Scientifique	P4.2 : Comparaison de la couverture et de la capacité d'annotation de plusieurs terminologies dans le domaine de l'anatomie (localisation des atteintes et des interventions de santé).	Cette étude évalue les performances de la SNOMED CT et de la CIM-11 sur le cas d'usage anatomie par rapport à d'autres terminologies spécifiques du domaine (Etude LIMICS).
Scientifique	P4.3 : Etude de la couverture relative de la SNOMED CT par rapport à une ontologie spécifique du parcours de soin de la maladie de Charcot (OntoParon).	Cette étude évalue les performances de la SNOMED CT sur un cas d'usage maladie rare (maladie de Charcot) (Etude LIMICS).
Scientifique	P4.4 : Comparaison des performances de la SNOMED CT et de la CIM-11 sur le cas d'usage oncologie.	Cette étude compare la performance de la SNOMED CT et de la CIM-11 sur le cas d'usage oncologie (Etude ISPED).
Scientifique	P4.5 : Evaluation de l'alignement entre la SNOMED CT et la CIM-11.	Cette étude évalue la capacité d'alignement et d'annotation de la SNOMED CT, de la CIM-11 et d'autres terminologies (Etude LIMICS).
Scientifique	P4.6 : Étude des besoins de terminologies et représentation des connaissances dans le cadre du RHU PsyCARE avec focus sur les –omiques - Analyse de la couverture relative de la SNOMED CT et d'autres terminologies par rapport à ces besoins.	Cette étude évalue la capacité de la SNOMED CT à couvrir un cas d'usage génomique (Etude LIMICS).
Scientifique	P4.7 : Comparaison des performances de la SNOMED CT à un modèle médicament (ROMEDI).	Ces deux études, conduites par le LIMICS et l'ISPED, évaluent la capacité de la SNOMED CT à couvrir un cas d'usage médicament.
Scientifique	P4.8 : Comparaison de la SNOMED CT à des ressources onto-terminologiques dans le domaine médicamenteux.	
Scientifique	P4.9 : Etude sur les besoins de codages et pour l'amélioration de la structuration des données en soins primaires.	Cette étude, conduite par le Pôle Médical et de Labellisation de l'ANS, se focalise sur l'activité en soins primaires, notamment le besoin de codage et l'amélioration de la structuration des données.
Scientifique	P4.10: Position Paper on Medical Knowledge Extraction via Word Embeddings.	Cette étude, présente un état des lieux effectué le laboratoire d'informatique de l'Ecole Polytechnique, des approches de pointe dans le domaine de l'extraction de connaissances.

6 ANNEXE : CARTOGRAPHIE DE LA SNOMED CT ET DES EQUIVALENCES TERMINOLOGIQUES

Le tableau ci-après liste les 19 différents chapitres de la SNOMED CT. Il identifie les terminologies couvrant également le domaine en usage en France ou en instruction *(terminologie en italique).* Les alternatives listées ne sont pas exhaustives.

A noter que certains chapitres sont transverses et sont susceptibles d'être utilisés dans de multiples cas d'usage.

Les chapitres grisés indiquent ceux-qui n'ont pas a priori d'usage immédiat en France.

Ce tableau a vocation à orienter les travaux d'instruction sur les futurs cas d'usage.

Un second tableau à la fin de l'annexe identifie des domaines pour lesquels la SNOMED CT n'est a priori pas pertinente.

Tableau 25 : Cartographie des domaines de connaissance de la SNOMED CT : Comparateurs possibles et cas d'usage en France (31/07/2020)

Chapitre SNOMED	Contenu ***Commentaire***	Terminologies couvrant le domaine en usage en France ***en instruction***	Domaines de Cas d'usage En France
Body structure 39 458 concepts	Ensemble des structures corporelles normales ou anormales (anatomie et histopathologie)	• CIM-11 (extension codes) • ADICAP en anatomo-cytopathologie Dictionnaires anatomie • *CIM-11, CIM-0 pour la description des tumeurs* • *Uberon (uber anatomy ontology), FMA (Foundational Model of Anatomy).* • *NCIT (national cancer institute thesaurus)*	• Cancérologie • ACP • Tout cas d'usage requérant une identification de site anatomique
Clinical findings 115 537 concepts	Observations cliniques (maladies et signes cliniques)	• CIM-11, CIM-10, CIM-O • CISP 2 • Nomenclature Orphanet des maladies rares • HPO (Signe cliniques) • *ICNP (international classification of nursing practice)* • *NCIT (national cancer institute thesaurus)* • *Clinical measurements ontology (medical college of Wisconsin)*	• PMSI, Maladies Chroniques (ALD), SNIIRAM, DMP, SNDS • Historique des soins (VSM) • SDM MR (maladie rares) • Lettre de liaison • Cancérologie • …
Environment or geographical location 1 836 concepts	Données administratives et géographique *Données génériques associées à des concepts très variés : unité grands brulés, Asie mineure…*	Nomenclatures et Bases administratives décrivant lieux et établissements • COG (INSEE) • FINESS (DREES) • NOS (ASIP Santé) • Nomenclatures SAE (DREES) • Annuaire des sites maladies rares (AP-HP) • *Terminologies du ROR* • ….	Tous les volets d'interopérabilité requérant des données administratives (services hospitaliers, nationalité)

(suite)

Chapitre SNOMED	Contenu *Commentaire*	Terminologies couvrant le domaine en usage en France *en instruction*	Domaines de Cas d'usage En France
Events 3 182	Codage d'événements (exclusion des interventions) *Ex : accident dû à un excès d'humidité, accident dû à une privation*	• CIM-10, CIM-11	Registre des décès
Observable entity 9 243	Evaluations en santé *Ex : listes de scores et échelles devant être post coordonnés pour y associer des résultats*	• LOINC • TNM • CIM-11 (section V) • *ICF (OMS)*	• CR biologie • FRCP, PPS cancer • Médicosocial handicap
Organism 35 072 concepts	Agents vivants importants en santé *Ex : bactéries, virus champignons, parasites*	• CIM-11 • LOINC • *NCBI Taxon (Classification du vivant)* • *NCIT (national cancer institute thesaurus)*	Bactériémies (AP-HP/HDH)
Pharmaceutical / biologic product (product) 22 502 concepts	Médicaments et produits biologiques *Ex : non spécifique du marché français, pas d''identification formelle* *Ex : Produits biologiques : Greffe produits sanguins.*	• Médicabase • BDPM (ANSM) • ATC (OMS) • Répertoire de médicaments orphelins (AP-HP) • Bases médicamenteuses du secteur marchand (Vidal BCB, Thésorimed, Thériaque, Clicadoc) • LOINC	Tout volet d'interopérabilité citant des médicaments /produits biologiques
(Suite)Physical force (physical force) 169 concepts	Forces physiques en relation avec la santé (irradiation, humidité, friction, etc.) *Ex : altitude, magnétisme, agent thermique ou irradiant*	Non investigué	Pas de cas d'usage a priori en France
Physical objects 15 725 concepts	Objets physiques *Le champ des concepts est très large (incluant dispositifs médicaux, train et avions)*	• CLADIMED • LPP (CNAM) • *CND/EMDN* • *Bases privées (Exhausmed, Phast, ACL, etc.)*	• Historique des soins • Circuit des dispositifs médicaux • E-prescription • Traçabilité des DMI
Procédure 58 290 concepts	Actes accomplis *Vaste périmètre du chapitre : ex : demande de passeport, plainte formelle contre l'hôpital, endoscopie, mesure de TA…*	• CCAM, NGAP, NABM (CNAM) • LOINC • CISP 2 • *ICHI (OMS)* • *Radlex play book (imagerie)* • *ICNP (actes infirmiers)*	• Historique des soins • Analyses biologiques • Imagerie

(suite)

Chapitre SNOMED	Contenu *Commentaire*	Terminologies couvrant le domaine en usage en France *en instruction*	Domaines de Cas d'usage En France
Qualifier value 11 022 concepts	Attributs de post coordination, *Ex : lettre grecque alpha, désignation de classifications, voie d'administration, action…*	Qualifiants répartis dans de multiples terminologies ou créés ad hoc au niveau de la syntaxe d'interopérabilité (e.g code system name et display Name pour les classifications)	NA
Record artifact 496 concepts	Eléments du dossier Identification de parties de dossiers patient *Ex : enregistrement de prescription médicamenteuse*	NA	NA
Situation with explicit context 4 862 concepts	Codage de constatations cliniques diverses (antécédents et événements futurs ou passés) *Ex : Conseil thérapeutique, évaluation des croyances spirituelles de la famille, mauvaise réponse au traitement.*	• CIM-10, CIM-11 • ICHI • LOINC • Vocabulaire HL7	Pas de cas d'usage spécifique Complément d'information en post coordination
SNOMED CT Model Component (metadata) 1 764 concepts	Chapitre technique Définitions des predicats et propriétés des ressources ontologiques de la SNOMED CT.	NA	Chapitre à usage interne pour la SNOMED CT
Social context (social concept) 4 434 concepts	Description situation sociale du patient	Nomenclatures de Catégorisation INSEE	Donnée administratives
Special concept 643 concepts	Termes génériques Ex : concept dupliqué résultats hématologiques limites…	*Incorporable dans des extensions nationales de la CIM ou recours à des terminologies lexicales de type Wordnet).*	Pas de cas d'usage spécifique
Specimen 1707 concepts	Description d'un l'échantillon	• ADICAP • LOINC (milieux) • ICHI • • *FMA (anatomie)*	• Volets ACP • CRGM • CR Biologie
Staging and scales 1 584 concepts	Stadifications diverses Catalogue des diverses possibilités de stadification **sans les résultats**	• CIM-11, CIM -O (codes extensions) • TNM	Pas de cas d'usage spécifique
Substance 26 853 concepts	Constituants chimiques	• LOINC (produits biologiques) • CIM-11 (allergènes, substances chimiques, enzymes, protéines diverses) • ATC (OMS) Active ingredients • BDPM (ANSM) composants	Description fine d'allergènes ou de médicaments

Il est à noter qu'au moins trois domaines de connaissance ne sont pas couverts par la SNOMED CT alors qu'un besoin existe en France (voir tableau ci-après)

Tableau 26 : Besoins non couverts par la SNOMED CT

Domaine de connaissance	Terminologies couvrant le domaine en usage en France *en instruction*	Domaines de Cas d'usage En France
Codage des gènes et protéines	• *HGNC (huogo gene nomenclature* • *GO (gene ontology)* • *NCBI Gene* • *Uniprot*	• Set de données minimum – maladies rares. • Compte-rendu de génétique moléculaire (cancérologie)
Codage des réactifs et techniques de laboratoires	• NRT (ANSM)	• Décret biologie 2016, mise en œuvre du contrôle national de qualité
Médico-social	• Seraphin SI-MDPH • *ICF OMS* • *ICHI (OMS)*	• SI MDPH