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

Annexe P4.1

Evaluation des terminologies du domaine microbiologique en contexte d'usage : Choix d'une terminologie de référence pour un catalogue d'agents infectieux à l'AP-HP

Décembre 2020

Classification : Restreinte *Version : Finale*

Contributeurs du document

Florent Desgrippes	Responsable de projets (ANS)
David Trystram	Responsable Pôle laboratoire (DSIP)
Sylvie Cormont	Biologiste expert sémantique (AP-HP)
Nicolas Griffon	Médecin de santé publique, expert interopérabilité (AP-HP)
Nicolas Paris	Responsable des développements de l'Entrepôt de Données de Santé (EDS) (AP-HP)
Christel Daniel	Directrice adjointe du domaine Données et Recherche (AP-HP)
Yann Briand	Pharmacien Expert Sémantique (ANS)

SOMMAIRE

1 INTRODUCTION

Entre 2014 et 2017, l'Agence du Numérique en Santé (ANS) a été missionnée par la Délégation ministérielle du Numérique en Santé (DNS) pour mener des travaux portant sur la mise en œuvre de référentiels sémantiques pour le secteur santé-social. Au cours de travaux successifs, l'ANS a montré l'importance d'une adoption de référentiels sémantiques facilement utilisables par les professionnels de santé afin de favoriser les échanges interprofessionnels et l'exploitation de bases de données structurées.

Plusieurs ressources sémantiques sont candidates. La SNOMED CT et les classifications de l'OMS ont notamment émergées pour différents cas d'usage. Le choix et le positionnement relatif de ces terminologies sont toujours débattus à ce jour.

Pour avancer le débat vers des conclusions objectives, la DNS, a saisi le Centre de Gestion des Terminologies de Santé (CGTS) de l'ANS pour mener une évaluation de ressources sémantiques comprenant la SNOMED CT (SCT) pour les positionner dans l'écosystème sémantique français. L'enjeu principal est de répondre à la question suivante : faut-il adopter la SNOMED CT ?

L'ANS a mené une étude apportant des éléments de réponse selon quatre axes :

- **Axe « benchmark » international** par la mise à jour des retours d'expérience internationaux sur la SNOMED CT effectués en 2017 lors de l'étude de phase IV[1] ;
- **Axe bibliographique** par la clarification des critères d'évaluation des ressources sémantiques et l'état de l'art de l'évaluation de la SNOMED CT ;
- **Axe juridique** par le positionnement relatif d'un ensemble de ressources sémantiques en termes de propriété intellectuelle et la clarification des conditions de mise en œuvre de la SNOMED CT en France ;
- **Axe scientifique** par l'étude du positionnement de la SNOMED CT dans plusieurs cas d'usage français.

Cette présente annexe correspond à un des cas d'usage des travaux scientifiques réalisés par le Pôle Médical et Labellisation (PML) de l'ANS, l'équipe WIND et le pôle laboratoire de l'AP-HP. Ce cas d'usage se focalise sur l'identification des bactéries dans le cadre de la création d'un référentiel de microorganismes par l'AP-HP.

2 CHOIX DU CAS D'USAGE

La prévention et le contrôle des infections bactériennes communautaires ou nosocomiales sont aujourd'hui une priorité majeure pour les systèmes de santé du monde entier. Notamment, il est important de suivre l'émergence d'organismes multi-résistants pour améliorer les actions et les plans visant à résoudre ce problème.

La création de réseaux de surveillance des infections bactériennes, nationaux et internationaux, est obligatoire pour les institutions de santé publique et les ministères de santé. En France, le recensement des souches d'agents biologiques est organisé par les centres nationaux de références.

Au niveau de la santé publique et l'épidémiologie, la surveillance de la résistance aux antibiotiques est coordonnée par Santé Publique France (SPF), le centre européen pour la prévention et le contrôle des

[1] https://esante.gouv.fr/sites/default/files/media_entity/documents/asip_termino_rapport_phase_4_vf1.3.1_vf.pdf

maladies (ECDC) auquel rapporte le réseau Européen de surveillance des microorganismes résistants (EARS-Net)[1]. La France participe à cette surveillance depuis 2001[2] pour des souches spécifiques (S. aureus et S. pneumoniae) puis a étendu sa participation pour d'autres souches (E. coli et Enterococcus, S pneumoniae, etc.).

Aujourd'hui, SPF et l'Observatoire National de l'Epidémiologie de la Résistance Bactérienne aux Antibiotiques (ONERBA) fournissent des données à l'EARS-Net.

L'ONERBA a mis en place une surveillance de proximité reposant sur des réseaux de laboratoire de biologies privés (ex : AFORCOPI-BIO), d'hôpitaux universitaires (ex : réseau Azay microbiologie[3]). Au niveau de la conservation des souches, l'Institut Pasteur a constitué une biobanque depuis 1892 : la collection de l'institut pasteur[4]. Elle contient en 2015 plus de 12000 souches bactériennes. Cette collection pérenne a pour vocation de s'enrichir par le biais de collaboration avec l'institut Pasteur et par le dépôt de souche par des chercheurs français ou internationaux. Elle a pour objectif la conservation des souches et la diffusion d'informations (propriétés, conservation, identification). La collection est gérée par le centre de ressource biologique de l'institut Pasteur (CRBIP) créé en en 2002 pour regrouper les différentes biobanques de l'Institut Pasteur.

Ce travail pointe le besoin de langage commun pour nommer les agents biologiques à des fins d'échanges de données entre systèmes d'information dans un contexte de soin, de recherche ou de veille sanitaire.

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 la base de données ainsi constituée afin de pouvoir la connecter avec des bases de données similaires dans d'autres hôpitaux. Elle souhaite aussi pouvoir rassembler les informations disponibles concernant les germes et l'évolution de leur résistance aux antibiotiques et les transmettre à l'ONERBA qui les comparera à celles obtenues dans les pays étrangers.

Dans cet objectif, l'AP-HP et les hôpitaux partenaires souhaitent adopter une terminologie de référence pour traiter leur besoin d'échange de données et d'exploitation de ces bases de données ; un langage commun compréhensible par toutes les parties prenantes facilitant l'exploitation et l'interopérabilité des données de microbiologie.

Un précédent travail mené en 2015[5] a identifié des terminologies de référence candidates pour constituer un langage commun en microbiologie. La SNOMED 3.5, la SNOMED CT et la LOINC avaient été identifiées comme ressources sémantiques en microbiologie. Depuis cette date, la SNOMED 3.5, qui est la version antérieure de la SNOMED CT, est tombée en obsolescence par absence de maintenance. La SNOMED CT est une terminologie multi-domaines, propriété de SNOMED International. La terminologie « Logical Observation Identifiers Names & Codes » (LOINC) est actuellement la référence internationale utilisée pour le codage des observations de laboratoire. Elle est publiée par le Regenstrief Institute. Ce travail avait aussi identifié la terminologie NCBI Taxonomy[6] comme ressource pour classer les parasites mais n'avait pas identifié les sections concernant les champignons et les bactéries. Cette étude va donc préciser et approfondir ces travaux par une étude systématique.

Le présent travail est un approfondissement actualisant la liste des terminologies de référence dans le domaine de la microbiologie. Il s'inscrit dans l'effort d'évaluation requis par l'étude demandée par la DNS.

Il 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.**

3 MATERIEL ET METHODE

La méthodologie adoptée consiste à :

- Identifier et qualifier les terminologies existantes pouvant servir de langage commun pour nommer les microorganismes et standardiser les bases de données hospitalières de microorganismes et en particulier celle de l'AP-HP ;
- Evaluer et comparer les terminologies candidates sur des critères lexicaux (couverture, structure, alignements) et d'organisation (relations entre concepts, relations entre bases) afin de choisir la plus pertinente.

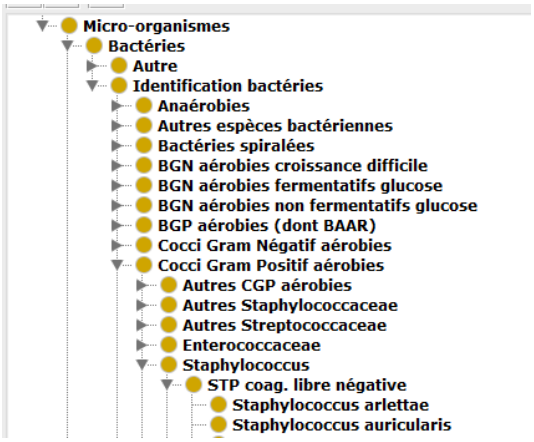
3.1 La base AP-HP

La base de données de microorganismes de l'AP-HP contient 687 champignons, 224 parasites, et 10 146 bactéries isolées chez l'Homme (base inspirée du Bergey's manuel et de la base Euzéby). Ce travail se focalise sur les bactéries. Celui des champignons et parasites se fera dans un second temps.

Concernant les 10 146 bactéries, il existe 4 niveaux de classements :

1) Le premier niveau décrit la forme de la bactérie (cocci, bacille, spiralées, intracellulaire), les caractéristiques de la paroi bactérienne (Gram positif ou gram négatif) et les caractéristiques de culture (aérobie ou anaérobie) ;
2) Le second niveau précise des groupes de bactéries ou des familles (mycobactéries, Enterococcacae...) ;
3) Le troisième niveau regroupe les genres (acinetobacter, streptococcus du groupe B…) ;
4) Le quatrième niveau détermine le genre et l'espèce (ex : « Terrabacter terrae »), ou le genre suffixé par « sp. » si l'espèce ne peut être déterminée (ex : « Terrabacter sp. »).

Figure 1 : La base AP-HP



3.2 Sélection des terminologies de référence candidates

Une liste de bactéries représentant diverses familles importantes de germes est établie à partir de la base de l'AP-HP.

Le tableau 1 présente les bactéries sélectionnées. Dix correspondent à des bactéries parmi les plus communes ou les plus étudiées sur le plan des résistances aux antibiotiques. Cinq autres correspondent à des bactéries peu communes présentes dans les sols ou les aliments humains.

Chaque bactérie est recherchée sur le serveur de terminologies Bioportal[b,7] pour identifier les terminologies dans lesquelles elles sont répertoriées.

L'objectif est la sensibilité de la cartographie des terminologies bactériennes afin d'avoir la liste la plus exhaustive des ressources sémantiques couvrant ce domaine de connaissance.

Tableau 1 : Echantillon de bactéries issues de la base AP-HP utilisées pour sélectionner les terminologies de référence

Bactéries	Espèces
Bactéries pour lesquelles des résistances sont connues	Acinetobacter baumannii
	Pseudomonas aeruginosa
	Staphylococcus aureus
	Mycobacterium tuberculosis
	Clostridium difficile
	Helicobacter pylori
	Campylobacter coli
	Klebsiella pneumoniae
	Vibrio parahaemolyticus
	Yersinia pestis
Bactéries peu communes	Carnobacterium alterfunditum
	Desulforhabdus amnigena
	Aeromicrobium panaciterrae
	Chainia fumigata
	Starkeya koreensis

Les terminologies dans lesquelles tout l'échantillon est présent sont :

1. auditées pour qualifier leur organisation ;
2. comparées à la base de données complète de l'AP-HP pour avoir une vision globale de leur couverture.

3.3 Audit des terminologies

L'évaluation des terminologies candidates a porté, d'une part sur des critères lexicaux et d'organisation des données et d'autre part, sur leur adéquation à couvrir le catalogue AP-HP de souches bactériennes.

3.3.1 Critères d'évaluation des terminologies

Trois catégories de critères sont investiguées :

1. **Critères lexicaux (**précision, couverture du domaine de connaissance, non ambiguïté, non redondance). Cet audit est utile pour une utilisation de la terminologie en interopérabilité.

2. **Critères relationnels (Classification et relations sémantique)** : organisation de l'arbre ou du réseau de concepts et termes. Cet audit est utile pour une utilisation de la terminologie en interopérabilité et en exploitation. **L'audit des relation externes** est utile quand la terminologie est utilisée par des chercheurs. En effet la mise en œuvre des principes de linked data est une solution efficace pour la représentation et la gestion de la connaissance[8] cet audit consiste à étudier les liens du concept (URI ou IRI) vers d'autres ressources du web (principe de « linked data ») afin d'apporter de nouvelles informations ou de lier les informations associées au concept.

3. **Accessibilité et exploitabilité : Propriété intellectuelle : examen des licences utilisateurs.**

[b] *Bioportal ((https:/bioportal.bioontology.org/)) est un serveur multi-terminologies développé par le Centre national d'ontologie biomédicale (NCBO). Le NCBO est l'un des centres nationaux d'informatique biomédicale. Il est financé par le NIH. L'un des objectifs du NCBO est de fournir des outils d'exploration et d'accès aux ontologies biomédicales, sous une forme exploitable par machine.*
Le NCBO est présent sur de multiples sites : Université de Stanford, Mayo Clinic, Université de Victoria et à Université de Buffalo.

3.3.2 Alignements sémantiques des terminologies étudiées avec le catalogue AP-HP : adéquation au besoin

Les 10 146 bactéries du jeu de valeur de l'AP-HP ont été comparées aux terminologies sélectionnées en employant une méthode de correspondance sémantique fondée sur les distances d'édition et la distance de Levenshtein en particulier.

La distance de Levenshtein[9] (d) est un calcul donnant une mesure de la différence entre deux chaînes de caractères. Elle est le nombre minimal de remplacements, ajouts et suppressions de caractères pour passer d'une chaine A à une chaîne B.

Plus la distance est faible, plus les chaînes de caractères sont similaires.

Cette méthode est utilisée notamment pour comparer des terminologies, produire des algorithmes d'analyse de séquence d'ADN (Acide Désoxyribonucléique) ou détecter des erreurs typographiques dans des textes[10,11,12,13].

Les concepts de la base AP-HP (genre espèce) sont en latin, langage de référence en microbiologie, ce qui facilite les alignements avec toute terminologie du domaine et rend la méthode peu sensible aux problématiques de traduction (français – anglais par exemple).

Des alignements automatiques sont proposés avec les terminologies de référence. Les propositions d'alignement associées aux distances sémantiques de Levenshtein les plus faibles sont retenues et ensuite analysées.

4 RESULTATS

4.1 Identification des terminologies de référence candidates

Le tableau 2 ci-après présente toutes les terminologies référencées sur Bioportal dans lesquelles tout ou partie l'échantillon des 15 bactéries a été retrouvé.

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.

Les 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).

Il est à noter que les bactéries se retrouvent également dans 18 autres terminologies montrant la richesse des ressources sémantiques dans ce domaine et le besoin de la communauté scientifique pour la structuration et le codage des microorganismes.

En raison de leur meilleure couverture a priori, NCBI Taxonomy et la SNOMED CT sont retenues comme terminologies candidates pour une analyse approfondie.

Les autres terminologies alternes sont détaillées en annexe de cette publication.

Tableau 2 : Echantillon de bactéries test - Terminologies contenant les bactéries de l'échantillon issues du catalogue AP-HP

Bactérie	NCBI Taxon	SNOMED CT	LOINC	NCIT	CIM-11	MeSH	Autres terminologies dans lesquelles les bactéries sont contenues
Acinetobacter baumannii	+	+	+	(*)	+	+	SNMI, IOBC, PLOSTHES, OCHV,
Pseudomonas aeruginosa	+	+	+	+	+	+	RxNORM, NDDF, IOBC, CRISP, SNMI, CSEO, PLOSTHES, OCHV, ONTOLUrgences, Fly taxonomy
Staphylococcus aureus	+	+	+	+	+	+	RxNORM, NDDF, IOBC, CRISP, SNMI, CSEO, PLOSTHES, GAMUTS, OCHV, ONTOLUrgences, ORTH, Fly taxonomy, AHOL
Mycobacterium tuberculosis	+	+	+	+	+	+	RxNORM, IOBC, CRISP, SNMI, CSEO, PLOSTHES, GAMUTS, ONTOLUrgences, ORTH, Fly taxonomy, OntoPNEUMO
Clostridium difficile	+	+	+	+	+	+	CRISP, SNMI, IOBC, CSEO, PLOSTHES, SYN, RGO, ONTOLUrgences
Helicobacter pylori	+	+	+	+	+	+	CPT, RxNORM, SNMI, IOBC, CSEO, PLOSTHES, OCHV, OntoaAD
Campylobacter coli	+	+	+	+	+	+	SNMI, IOBC, CSEO, CRISP
Klebsiella pneumoniae	+	+	+	+	+	+	RxNORM, NDDF, IOBC, CRISP, SNMI, CSEO, PLOSTHES, OCHV, ONTOLUrgences,
Vibrio parahaemolyticus	+	+	+	+	+	+	SNMI, CSEO, IOBC, OCHV
Yersinia pestis	+	+	+	+	+	+	RxNORM, CRISP, SNMI, CSEO, PLOSTHES, OCHV, ONTOLUrgences, ORTH
Carnobacterium alterfunditum	+	+	non	non	non	+	IOBC, Fly taxonomy
Desulforhabdus amnigena	+	+	non	+	non	non	-
Aeromicrobium panaciterrae	+	(**)	non	non	non	non	-
Chainia fumigata	(***)	+	non	non	non	non	-
Starkeya koreensis	+	(**)	non	non	non	non	-
Total genre espèce	15/15	13/15	10/15	11/15	10/15	11/15	11/15 en considérant tout le panel de terminologies

** Présent sous forme de synonyme (acinetobacter anitratus) dans le NCIT*
*** présent au niveau du genre : genre Aeromicrobium ou genre starkeya. Les espèces sont non répertoriées*
**** Présent sous forme de synonyme (Streptomyces fumigatiscleroticus) dans NCBI Taxonomy*
(4) Le tableau mentionne "+" si la bactérie donne au moins 1 résultat de recherche dans le moteur de recherche du site bacterio.net et "+" si la bactérie est présente dans la liste des procaryotes connus de bacterio.net disponible aux URL : http://www.bacterio.net/-allnamesac.html , http://www.bacterio.net/-allnamesdl.html , http://www.bacterio.net/-allnamesmr.html , http://www.bacterio.net/-allnamessz.html

4.2 Audit des terminologies

4.2.1 Critères lexicaux

4.2.1.1 Termes et concepts

4.2.1.1.1 SNOMED CT

La SNOMED CT[xiv] a été créée à l'origine par le College of American Pathologists. Elle est actuellement gérée par SNOMED International. La SNOMED CT fournit un ensemble très complet et détaillé de termes cliniques utilisés par de nombreux système d'information pour structurer les informations des dossiers de santé électroniques (eHR pour "electronic health record" en anglais). La version de SNOMED CT de juillet 2019 contient 350 829 concepts actifs.

La SNOMED CT est une terminologie organisée en 19 chapitres. Le chapitre "organismes" ("Organism (organism)", SCTID : 410607006) contient 34 773 concepts dont 13 269 sont liés aux bactéries ("Domain Bacteria (organism)", SCTID : 409822003). 11 508 bactéries sont définies au niveau de l'espèce. Les critères taxonomiques peuvent être associés à d'autres critères pré-coordonnés (sérogroupes, résistance à un antibiotique...) (cf. tableau 3).

La SNOMED CT n'est pas une source officielle pour les microorganismes. Elle compile la connaissance issue d'autres sources.

Les concepts sont identifiés par un SCTID et par URIs (ex : 413422008 ou http://snomed.info/id/413422008 pour "Achromobacter spanius (organism)").

Ce travail est basé sur la version 2017 et du 31/07/2019 de SNOMED CT[3].

4.2.1.1.2 NCBI Taxonomy

Le projet de constitution de NCBI Taxonomy remonte à 1991. La base a été développée par le National Center for Biotechnology Information (NCBI/USA) pour fournir une nomenclature pour tous les organismes présents dans la base de données de l'international Nucleotide sequence database collaboration (INSDC). Cette base regroupe notamment les banques de gènes européennes (ENA) et japonaise (DDBJ).

NCBI Taxonomy est une terminologie du vivant servant de passerelle vers d'autres terminologies publiées par le NCBI. En 2011, elle contenait déjà toutes les espèces de procaryotes (archées et bactéries) formellement décrites.

Néanmoins elle n'est pas une source officielle de taxonomie des bactéries mais elle s'appuie sur les travaux de grands projets de classification du vivant pour répertorier les séquences génomiques au sein de la base INSDC[4] : Catalog of life, Encyclopedia of life, NameBank, wikiSpecies, etc.

NCBI Taxonomy contient 1 776 639 concepts dont 375 961 concepts liés aux bactéries parmi lesquels 371 766 bactéries définies au niveau de l'espèce (version de juillet 2018) (cf. tableau 3).

NCBI Taxonomy est accessible par plusieurs canaux : site web de Bioportal, site du NCBI[5] ou plateforme ontobee[6].

[3] http://www.snomed.org/snomed-ct/get-snomed
[4] International Nucleotide Sequence Database Collaboration (INSDC)
[5] Serveur ftp : ftp://ftp.ncbi.nih.gov/pub/taxonomy, browser : https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi
[6] http://www.ontobee.org/ontology/NCBITaxon. Ontobee est le serveur de données liées par défaut pour la publication et la navigation dans les ontologies biomédicales de la bibliothèque Open Biological Ontology (OBO) Foundry (http://obofoundry.org). Ontobee héberge actuellement plus de 180 ontologies (dont 131 ontologies OBO Foundry Library) avec plus de quatre millions de termes.

Les concepts sont identifiés par URIs. Staphylococcus Aureus possède plusieurs URIs en fonction du canal de distribution mais est toujours identifié par le même code taxon 1280 :

- http://purl.bioontology.org/ontology/NCBITAXON/1280 sur le portail Bioportal ;
- http://purl.obolibrary.org/obo/NCBITaxon_1280 sur la plateforme ontobee.

4.2.1.2 Couverture et précision

Le tableau 3 compare les couvertures respectives de SNOMED CT et NCBI Taxonomy.

Les arbres taxonomiques de SNOMED CT et NCBI Taxonomy présentent des différences notables au niveau quantitatif. **NCBI Taxonomy est plus complète que SNOMED CT :**

- 4195 concepts de classification dans NCBI Taxonomy contre 1761 dans SNOMED CT ;
- 371 766 espèces de bactéries dans NCBI Taxonomy contre 11 508 dans SNOMED CT.

Dans NCBI Taxonomy, 91 442 espèces sont formellement décrites. Les 280 324 autres espèces sont marquées en cours de description et d'identification : genre associé au suffixe "Sp.".

Dans SNOMED CT, 6391 espèces de bactéries sont formellement décrites. 5117 sont décrites avec une ligne éditoriale propre à la SNOMED CT :

- Il existe ainsi 2707 variétés de Salmonelles décrites avec genre et nom de ville (Salmonella Duisburg, Duval, ealing, Dublin…) ou associée à des sérogroupes ("Salmonella enterica subspecies enterica serovar 6,7: -:e,n,z15" - SCTID: 442115007 - ou "Salmonella IIIb 60:r:z" - SCTID: 398353003) ;
- On retrouve ainsi 362 streptocoques, 667 eschercica suivant des lignes éditoriales similaires ;
- Des bactéries sont préfixées par le CDC (ex : "Centers for Disease Control and Prevention Corynebacterium group A-3" - SCTID : 243268004).

A chaque strate de la classification, on peut constater de manière générale qu'il existe plus de phylum, classe, ordre, famille et genre dans NCBI Taxonomy que dans SNOMED CT (tableau 3).

Les différences au niveau des {ordre, super phylum, sous phylum, sous ordre, sous classe} peuvent résulter de choix éditoriaux plus que de différences de couverture.

Ainsi, au niveau super phylum, 3 super phylum semblent manquants dans NCBI Taxonomy. Les embranchements sont en fait représentés différemment dans les deux terminologies :

- Le super phylum "Bacteroidetes/Chlorobi group" (SCTID : 415672001) de SNOMED CT est représenté dans un groupe de phylum au sein de NCBI Taxonomy : le FCB group (Fibrobacteres, Chlorobi, and Bacteroidetes) qui n'a pas de classification spécifique.

- Le super phylum "Fibrobacteres/Acidobacteria group" (SCTID : 426797003) de SNOMED CT n'est pas référencé dans NCBI Taxonomy. Dans cette terminologie, le phylum fibrobacteres est apparenté au groupe FCB décrit dans la littérature[xv].
  Par ailleurs, ce super phylum n'a pas de descendance du côté acidobacteria dans SNOMED CT rendant l'arborescence incomplète. Seul le phylum fibrobacteres est décliné en arborescence.

- Dans NCBI Taxonomy, ce phylum fibrobacteres est décliné à partir du FCB group en parallèle avec les phylums Chlorobi et bacteroides.

- De même, le super phylum "Chlamydiae/Verrucomicrobia group" (SCTID : 427433006) de SNOMED CT n'est pas représenté dans NCBI Taxonomy mais est inclus au sein du groupe PVC (Planctomycetes, Verucomicrobia, Chlamydiae) décrit dans la littérature[xvi]. Comme pour le super phylum précédent, la branche "Chlamydiae" n'a pas de déclinaison dans SNOMED CT rendant la définition de son concept parent ("super phylum **Chlamydiae**/Verrucomicrobia") inconsistante.

Tableau 3 : Comparaison des couvertures respective de NCBI Taxonomy et SNOMED CT dans le domaine bactérien

Classe taxonomique		SNOMED CT	NCBI Taxonomy	Différence	
				(NCBI Taxonomy -SCT)	%
Super phylum		*3*		-3	-
Phylum		*26*	*143*	117	450%
Subphylum			*1*	1	-
Class		*38*	*82*	44	116%
SubClass		*9*	*3*	-6	-67%
Order		*91*	*197*	106	116%
SubOrder		*17*	*8*	-9	-53%
Family		*245*	*460*	215	88%
SubFamily		*1*	*1*	0	0%
Tribe		*3*	*2*	-1	-33%
Genus		*1326*	*3221*	1895	143%
Sub Genus		*2*	*1*	-1	-50%
Species group		***	*70*	70	-
Species subgroup		***	*6*	6	-
Total concepts de classification		**1761**	**4195**		
Espèces	**Espèces définies formellement***	*6391*	*91 442*		
	Espèces définies hors taxonomie ou en cours d'identification**	*5117*	*280 324*		
	Total concept	**11508**	**371 766**		

Total concepts taxonomiques	**8 152**	**375 961**
Total concepts hors taxonomie	**5 117**	**-**

** Ex : Acetobacter orleanensis*
*** Ex : Pseudomonas sp. 5KW-VPa pour une espèce décrite mais pas formellement identifiée, Carbapenem resistant Klebsiella pneumoniae, Streptococcus pneumoniae Danish serotype 28F*
****des sous-groupes d'espèces sont présentes dans la SNOMED CT identifiées dans le libellé du concept. Ils ne sont pas identifiés au niveau de l'arborescence.*

L'exemple du genre Helicobacter apporte des détails qualitatifs expliquant les différences de couverture entre NCBI Taxonomy et SNOMED CT.

Pour référence, le catalogue AP-HP est composé de 33 bactéries du genre Helicobacter.

NCBI Taxonomy répertorie 59 bactéries du genre Helicobacter et en liste plus 161 de plus en cours d'identification dans le monde (nommage de type "Helicobacter sp. XXXXX").

La SNOMED CT liste 24 bactéries du genre Helicobacter.

Les 33 bactéries du genre Helicobacter du catalogue AP-HP ont toutes leur correspondance dans NCBI Taxonomy (cf. tableau 3). Par contre 9 d'entre elles n'ont pas de correspondance dans SNOMED CT.

Pour le genre Helicobacter, NCBI Taxonomy couvre donc 100 % du catalogue AP-HP alors que SNOMED CT n'en couvre que 73%.

4.2.1.3 Synonymes

NCBI Taxonomy et SNOMED CT ont des politiques éditoriales différentes. Il semble que SNOMED CT n'utilise que très peu les synonymes par rapport à NCBI Taxonomy.

Le tableau 5 présente quelques exemples pour Nesseiria Meningitidis, Bacteroides Fragilis et Acetobacter Aceti.

NCBI Taxonomy répertorie entre 5 et 15 « synonymes » et « désignations alternatives » pour chaque bactérie, alors que SNOMED CT ne donne pas de synonyme (Acetobacter Aceti) ou se limite à un (Bacteroides Fragilis) ou deux (Neisseria Meningitidis).

A noter que les synonymes NCBI Taxonomy ne sont pas présents dans le catalogue SNOMED CT et que parmi les synonymes SNOMED CT figurent des noms usuels (de type "Méningococcus") non présents dans une classification taxonomique.

Cette analyse serait à approfondir par une approche systématique permettant de conclure sur l'ensemble des terminologies

Tableau 4 : Correspondances base AP-HP, NCBI Taxonomy, SNOMED CT pour le genre Helicobacter

Catalogue AP-HP	Présence dans NCBI Taxonomy	Présence dans SNOMED CT
Helicobacter acinonychis	+	+
Helicobacter aurati	+	+
Helicobacter bilis	+	+
Helicobacter canadensis	+	+
Helicobacter canis	+	+
Helicobacter cholecystus	+	+
Helicobacter cinaedi	+	+
Helicobacter felis	+	+
Helicobacter fennelliae	+	+
Helicobacter ganmani	+	+
Helicobacter heilmannii	+	+
Helicobacter hepaticus	+	+
Helicobacter mesocricetorum	+	+
Helicobacter muridarum	+	+
Helicobacter mustelae	+	+
Helicobacter nemestrinae	+	+
Helicobacter pametensis	+	+
Helicobacter pullorum	+	+
Helicobacter pylori	+	+
Helicobacter rappini (nv: Flexibacter canadensis)	+	+
Helicobacter rodentium	+	+
Helicobacter salomonis	+	+
Helicobacter trogontum	+	+
Helicobacter typhlonius	+	+
Helicobacter anseris	+	non
Helicobacter bizzozeronii	+	non
Helicobacter brantae	+	non
Helicobacter cetorum	+	non
Helicobacter cynogastricus	+	non
Helicobacter equorum	+	non
Helicobacter marmotae	+	non
Helicobacter mastomyrinus	+	non
Helicobacter sp.	+	non
	+ H. ailurogastricus ,H. apodemus H. apri, H. baculiformis, H. callitrichis H. canicola H. cf. canis 29176 H. cf. pullorum, H. genosp. FL56 H. himalayensis,H. jaachi, H. macacae, H. magdeburgensis,H. muricola, H. peregrinus,H. saguini, H. suis H. suncus, H. tursiopsae H. valdiviensis H. vulpecula, H. winghamensis, Helicobacteraceae bacterium 4484_230 Helicobacteraceae bacterium CG1_02_36_14, Helicobacteraceae bacterium CG2_30_36_10, Helicobacteraceae bacterium UBA6016	
33 bactéries	59 bactéries identifiées + 161 bactéries de genre helicobacter en cours de définition	24 bactéries
	Taux de couverture > 100 %	**Taux de couverture = 73%**

Tableau 5 : Nommage des bactéries dans NCBI Taxonomy et SNOMED CT (exemples de Neisseiria Meningitidis, Bacteroides Fragilis et Acetobacter Aceti)

	NCBI Taxonomy	SNOMED CT
Neisseiria menigitidis	Synonyme : Neisseria weichselbaumii Alt label Diplokokkus intracellularis meningitidis Micrococcus meningitidis cerebrospinalis Micrococcus meningitidis Micrococcus intracellularis	Alt label Meningococcus Neisseria intracellularis
Bacteroides fragilis	Synonyme : Ristella fragilis Alt Label 1Sphaerophorus inaequalis 2Bacteroides incommunis 3Ristella uncata 4Pseudobacterium fragilis 5Ristella incommunis 6Bacteroides inaequalis 7Pseudobacterium uncatum 8Bacteroides uncatus 9Pseudobacterium incommunis 10Fusiformis fragilis 11Bacillus fragilis 12Pseudobacterium inaequalis 13Sphaerophorus intermedius	Alt label Bacteroides fragilis ss. fragilis
Acetobacter Aceti	Synonyme : Acetobacter aceti subsp. aceti Alt label 1Bacterium acetigenoidum 2Bacterium aceti 3Acetobacter lafarianum 4Acetimonas aceti 5Bacteriopsis aceti 6Micrococcus aceti 7Acetobacter ketogenum 8Acetobacter aceti orleanensis 9Mycoderma aceti souches non visqueuses (membraneuses) 10Bacterium hansenianum 11Acetobacter (subgen. Acetobacter) aceti 12Acetobacter aceti aceti 13Acetobacter aceti var. muciparum 14Bacillus aceticus	Pas de synonymes

4.2.2 Critères relationnels

4.2.2.1 Classification et relations hiérarchiques dans la SNOMED CT et dans NCBI Taxonomy

La classification des bactéries au sein de la SNOMED CT est multiaxiale.

Il existe une classification taxonomique qui remonte de l'espèce au domaine ("kingdom") bactérien : espèce/genre/famille/ordre/classe/phylum/kingdom.

En parallèle, il existe d'autres liens associés à des critères de description (forme, propriétés enzymatiques, caractéristique de la paroi bactérienne (coloration gram)). La figure 2 donne deux exemples de représentation de bactéries ("Yersinia Pestis" (SCTID : 54365000) et "Staphylococcus Aureus" (SCTID : 3092008).

SNOMED CT utilise un chemin logique pour représenter les connaissances terminologiques. Par exemple, le terme "Catalase-positive Gram-positive coccus" (SCTID : 243226004) inclut le fait que la bactérie appartient à la classe des cocci, et il implique également qu'il est "gram positif" ; la propriété "catalase positive" n'est pas retenue dans le chemin logique. On retrouve ainsi les deux concepts "cocci" et "bactérie gram positive" en tant que concepts parents (cf. figure 3) mais le concept "catalase positive" ne lui est pas apparenté.

Toutefois la comparaison de la représentation de Yersinia Pestis (bacille gram négatif responsable de la peste) et de Staphylococcus Aureus (cocci gram positif responsable d'infection de diverses localisations) révèle des différences de modélisation entre ces deux bactéries au sein de la SNOMED CT.

Plusieurs points de différences sont à noter :

- Les distances par rapport au nœud du domaine bactérien ne sont pas comparables :
  - Pour Yersinia Pestis, le phylum se situe au 5eme niveau sous le domaine bacterium ;
  - Pour Staphylococcus Aureus, le phylum se situe au 3eme niveau ;
- Les chemins sont différents : le phylum de Yersinia Pestis se trouve sous le nœud "Bacterium (organism)" alors que le phylum de "Staphylococcus Aureus" se trouve en parallèle du même nœud sans lien de parenté avec lui ;
- L'arbre taxonomique de Staphylococcus Aureus est **interrompu entre genre et espèce pour intercaler un nœud de caractéristique enzymatique**, le rendant non comparable à celui de Yersinia Pestis.

La classification des bactéries au sein de de NCBI Taxonomy est purement taxonomique.

La figure 4 présente la représentation de Staphylococcus Aureus et Yersinia Pestis rendant celle-ci très lisible

La comparaison NCBI Taxonomy vs SNOMED CT permet de visualiser les échelons de classification omis au niveau de SNOMED CT, ainsi que ceux rajoutés. SNOMED CT simplifie la représentation taxonomique et intercale des caractéristiques constitutives ou biochimiques.

La description des bactéries n'est pas uniforme au sein de SNOMED CT ce qui Limite les possibilités de classifications automatiques.

La figure 5 présente la comparaison des arbres de deux bactéries cousines appartenant à l'ordre des burkholderiales : Sutterulla stercrnicanis et Thiobacter subterraneous. Ainsi la définition de la première est enrichie par des caractéristiques morphologiques (bacille)[7] alors qu'une telle information n'est pas disponible pour la deuxième, bien qu'elle soit aussi un bâtonnet.

Ceci montre la difficulté de maintenir une description uniforme pour tous les concepts d'une terminologie de plus de 350 000 concepts. Les informations manquantes limitent les capacités de raisonnement pour une machine.

L'emploi de SNOMED CT en raisonnement doit être précédé d'un processus de mise en forme des définitions/relations pour permettre l'utilisation de la SNOMED CT sur ces cas d'usage.

[7] Bactérie allongée de forme bâtonnet.

Figure 2 : Arborescence de la SNOMED CT pour Yersinia Pestis

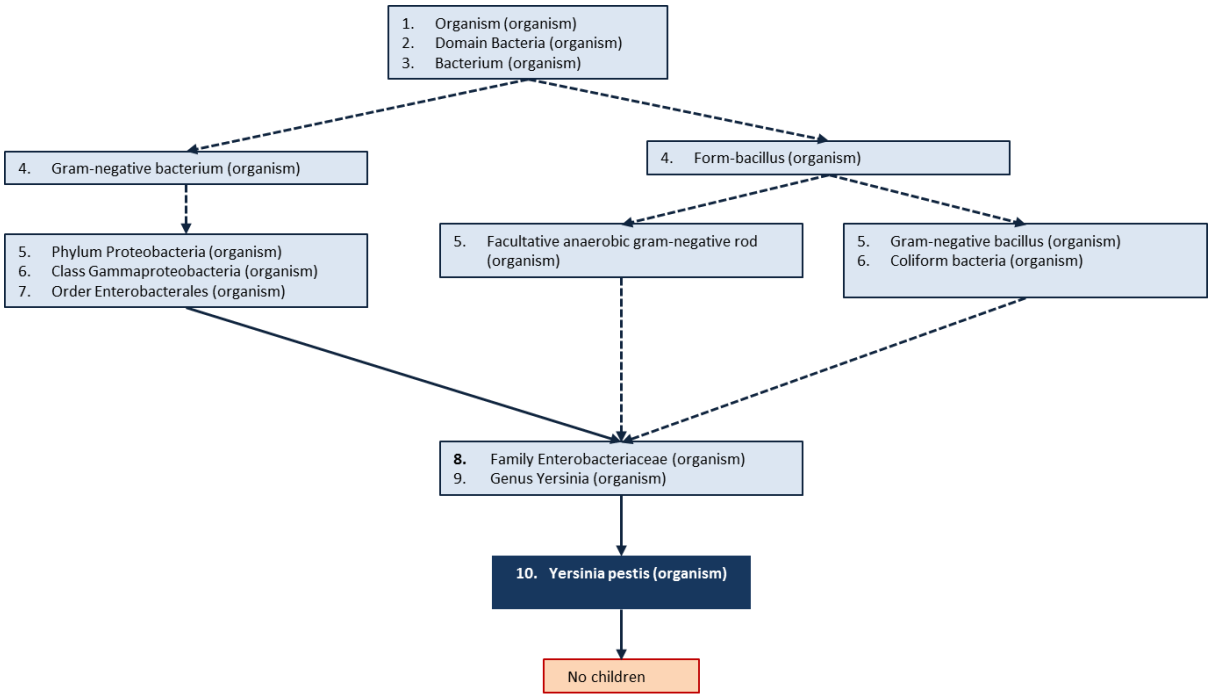
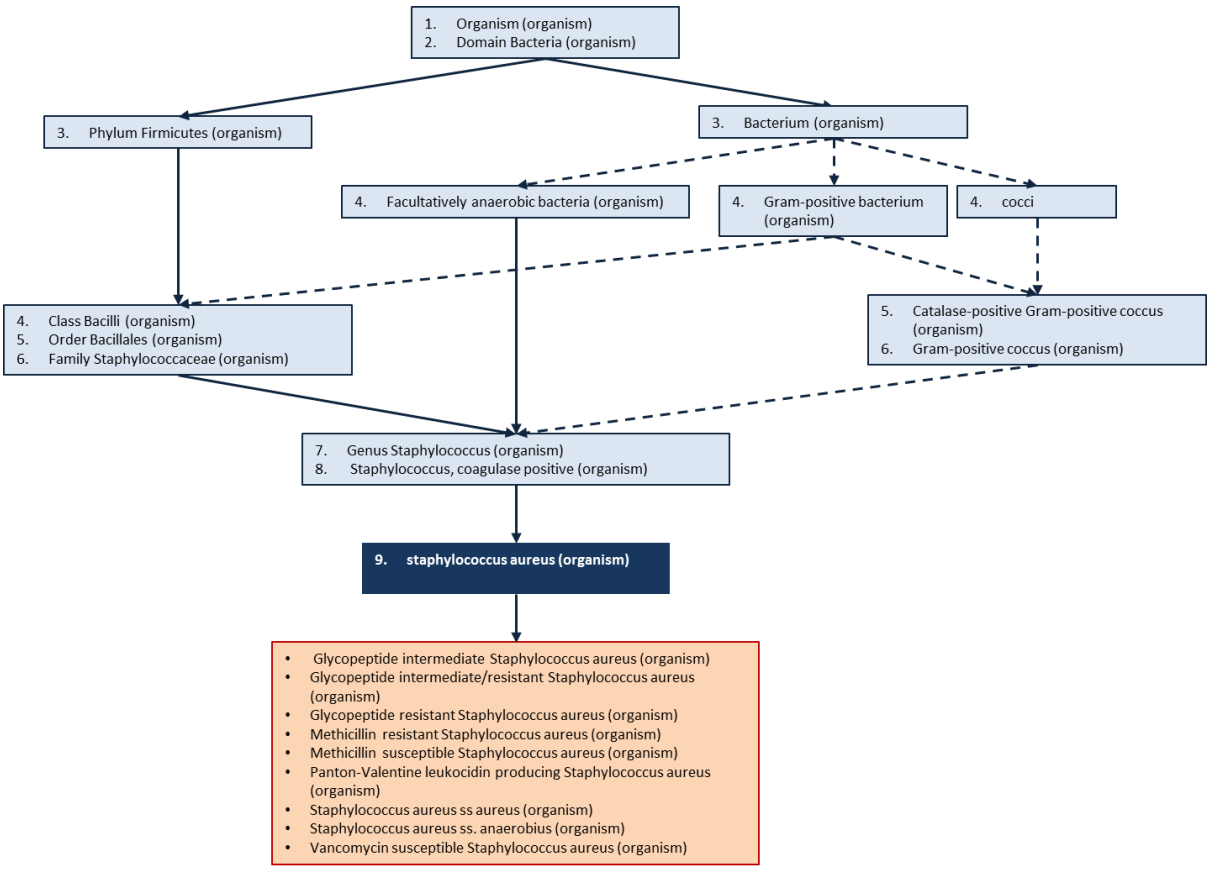


Figure 3 : Arborescence de la SNOMED CT pour Staphylococcus Aureus



NB : la hiérarchie taxonomique est figurée en trait plein.

Figure 4 : Arborescence de NCBI Taxonomy pour Yersinia Pestis et pour Staphylococcus Aureus

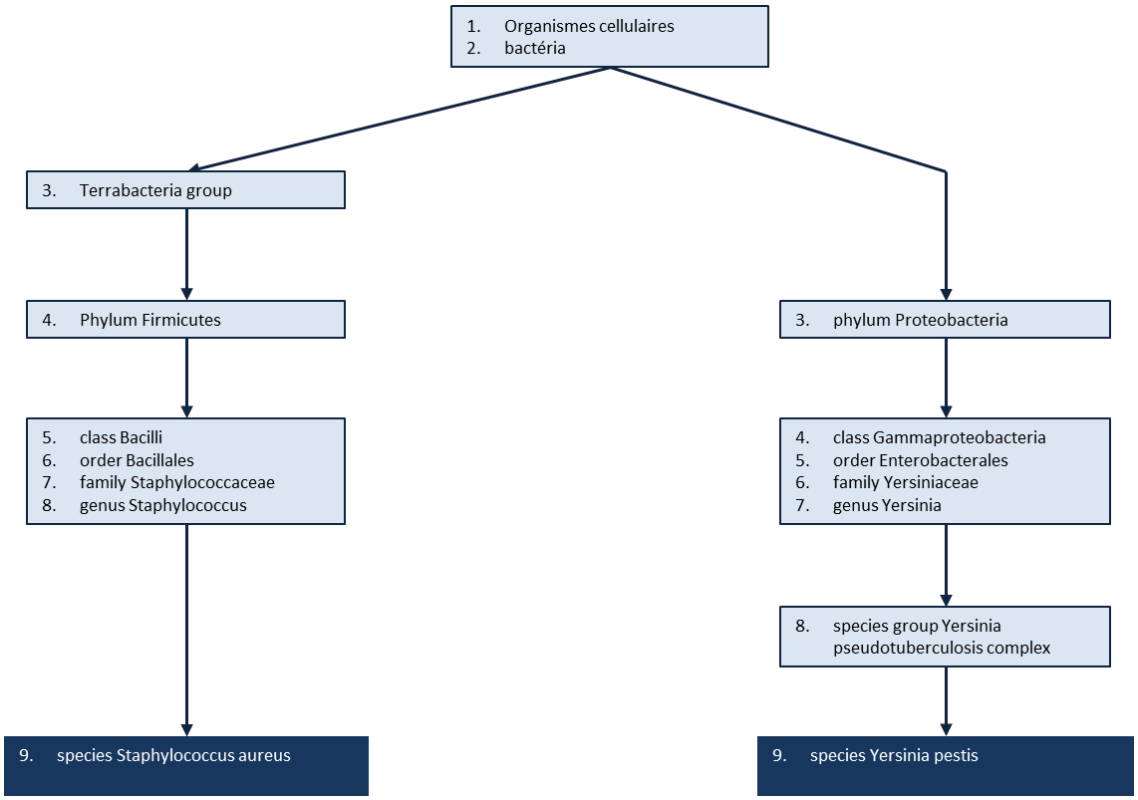
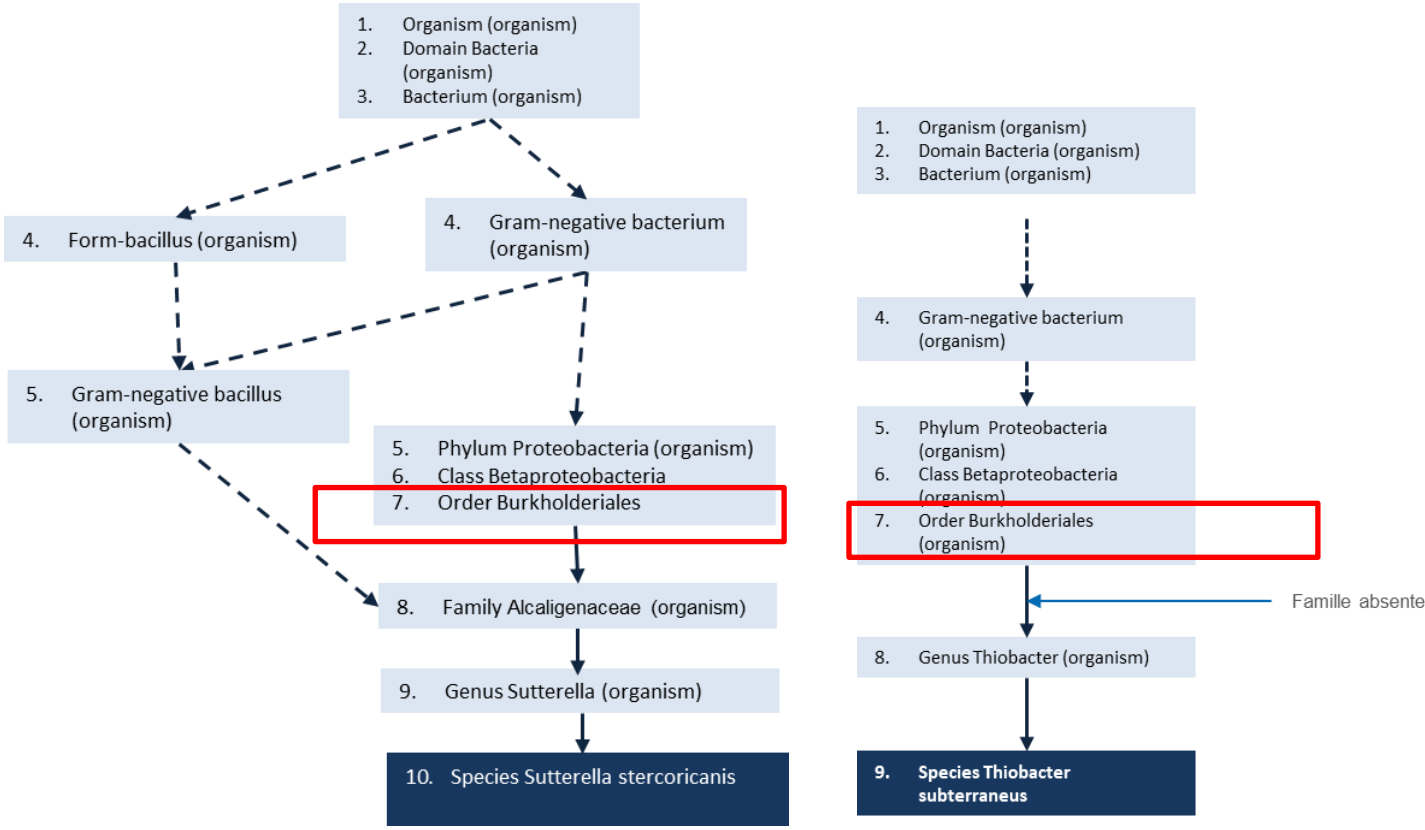


Figure 5 : Représentation des bactéries d'ordre des Burkolderiales
Exemple de Sutterella Strecoricanis et de Thiobacter Subterraneus



4.2.2.2 Données liées

4.2.2.2.1 SNOMED CT

Les URIs des concepts SNOMED CT se résolvent dans un navigateur web. Elles affichent la page web du navigateur SNOMED CT sans autres liens vers des ressources externes.

Les données de SNOMED CT ne sont pas des données liées.

4.2.2.2.2 NCBI Taxonomy

NCBI Taxonomy, via l'URI du concept placé dans un navigateur internet, donne accès à de multiples ressources quel que soit son canal de diffusion (NCBI, Bioportal, ou Ontobee).
Bioportal et Ontobee offrent les services d'un serveur multi-terminologies : cartographie des alignements et accès aux ontologies sur lesquelles les concepts sont alignés.

Le code taxon (ex taxid :1773 'identifiant du mycobacterium tuberculosis) est également un identifiant reconnu dans plusieurs « repository » documentaires, tels que Bacdive[8] (Bacterial Diversity metadatabase), Scope[9] (structural classifications of proteins), et

NCBI en tant que base de connaissance, donne accès à toutes les ressources bibliographiques à sa disposition :
- Références bibliographiques sur Pubmed ;
- Séquençage génomique ;
- Caractérisation des protéines bactériennes ;
- Bibliothèque des essais biologiques ;
- Code de la souche dans les diverses biobanques mondiales ;
- ….

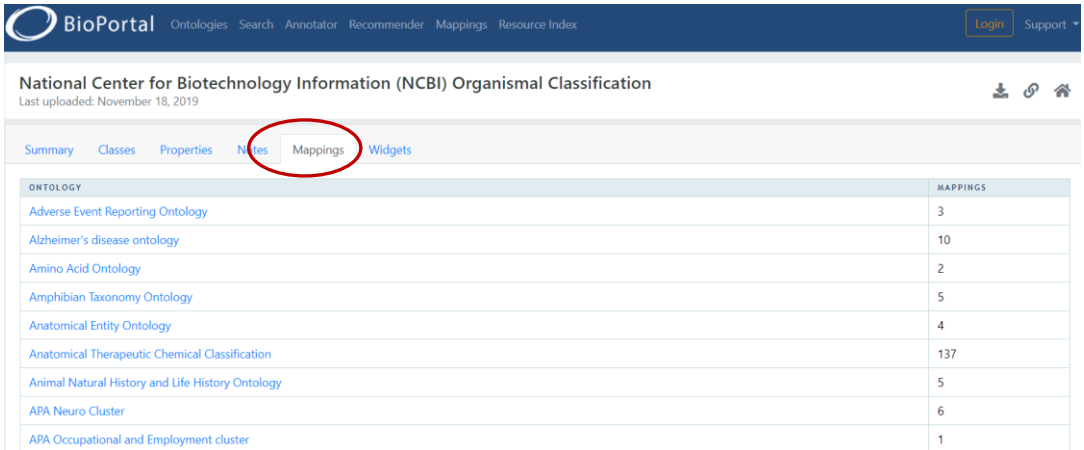
NCBI propose aussi un service de liens externes « LinkOut » (cf. figure 7)[10]. Tout utilisateur peut devenir un fournisseur de lien externe et ainsi enrichir la base de connaissance.

Les figures 6 et 8 illustrent la possibilité de liens de données avec Bioportal et Ontobee.

Les données de NCBI Taxonomy sont liées et interagissent avec de multiples sources de données, générant une grande ouverture vers la connaissance, au-delà à de la simple interopérabilité de la donnée.

Figure 6 : NCBI Taxonomy : données liées : accès à des mapping et sources de connaissances via les canaux d'accès à NCBI Taxonomy (Bioportal)

Bioportal : *http://bioportal.bioontology.org/ontologies/NCBITAXON?p=mappings*



[8] https://bacdive.dsmz.de/
[9] https://scop.berkeley.edu/sunid=89058
[10] https://www.ncbi.nlm.nih.gov/projects/linkout/doc/linkout.html

Figure 7: NCBI Taxonomy : données liées : accès à des mapping et sources de connaissances via les canaux d'accès à NCBI Taxonomy (NCBI)

NCBI : https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi

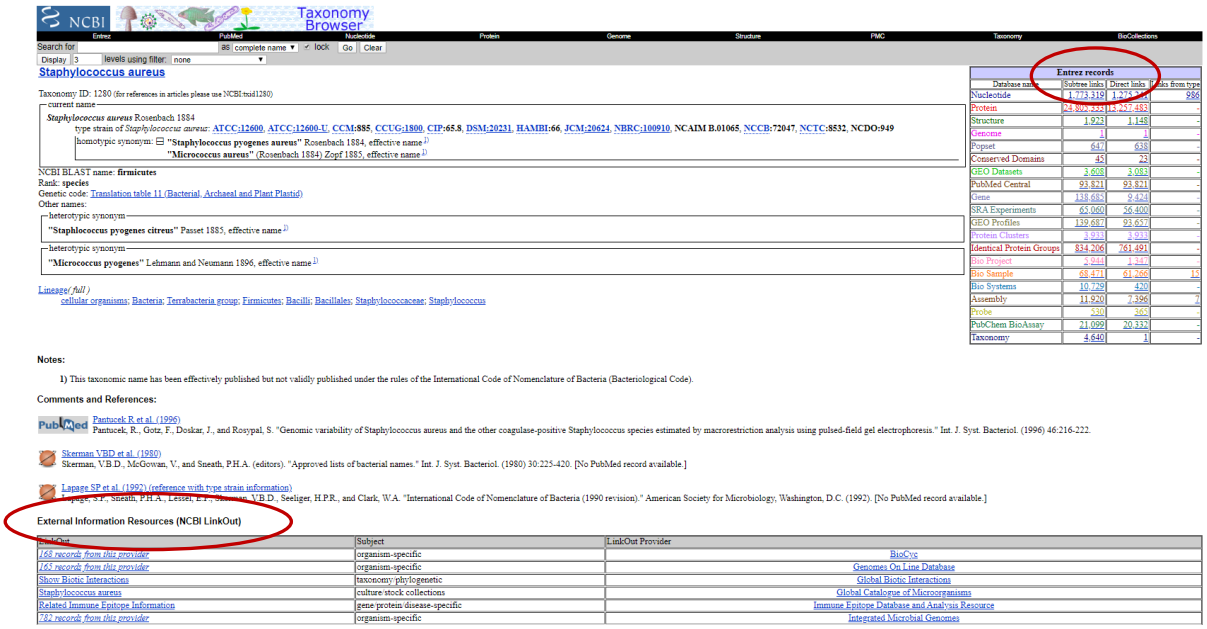
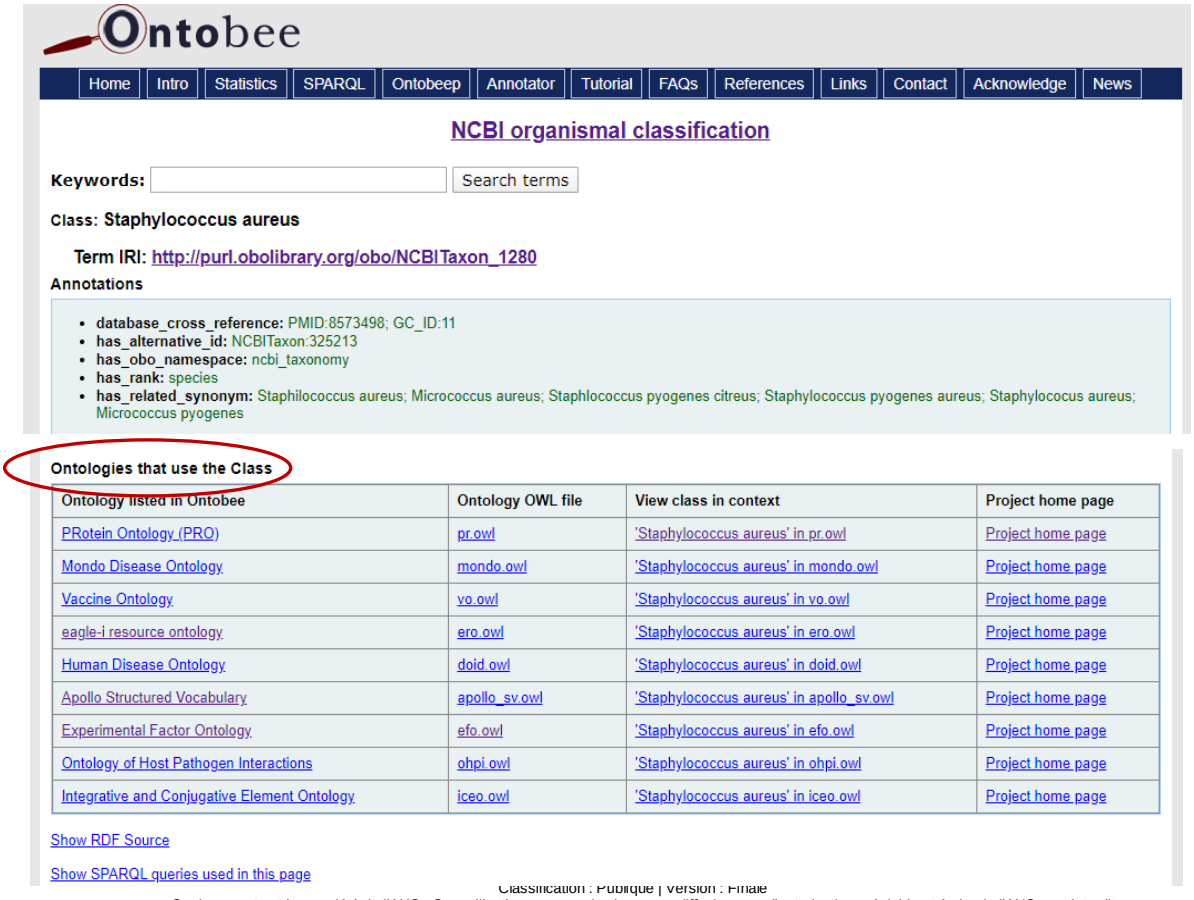


Figure 8 : NCBI Taxonomy : données liées : accès à des mapping et sources de connaissances via les canaux d'accès à NCBI Taxonomy (ontobee)

ontobee : http://www.ontobee.org/ontology/NCBITaxon?iri=http://purl.obolibrary.org/obo/NCBITaxon_1280



4.2.3 Propriété intellectuelle

4.2.3.1 SNOMED CT

SNOMED CT est une terminologie propriétaire. L'utilisateur doit prendre une licence affiliée[11] payante. Cette licence est gratuite si le pays dans lequel l'utilisateur met en œuvre des applications embarquant la SNOMED CT a pris une licence nationale. Les détails de de l'analyse juridique d'une adhésion à SNOMED international sont donnés en **annexe P2.0**.

4.2.3.2 NCBI Taxonomy

NCBI Taxonomy est une terminologie ouverte disponible gratuitement sous licence du domaine public CC0 1.0 universal[12] (voir analyse en annexe P2.0).

4.2.4 Alignements sémantiques entre la base AP-HP, la SNOMED CT et NCBI Taxonomy

4.2.4.1 Analyse quantitative

Des correspondances exactes (distance = 0) de la base AP-HP ont pu être établies avec respectivement 8040 et 5900 concepts bactéries (genre-espèce) de NCBI Taxonomy et de la SNOMED CT (cf. tableau 6).

Au total, respectivement 79.8% et 58.1% du catalogue AP-HP est directement aligné avec NCBI Taxonomy et SNOMED CT (étiquettes principales ou synonymes).

Les propositions d'alignement ont des distances moyennes de 3.2 (NCBI Taxonomy) et 4.4 (SNOMED CT) par rapport aux concepts répertoriés dans la base de l'AP-HP (cf. tableau 6).

1127 concepts de la base AP-HP sont alignés avec NCBI Taxonomy avec des distances comprises entre 1 et 20 (864 + 117 + 146). Ils sont de 3297 (2088 + 984 + 225) pour la SNOMED CT.

Au total, le meilleur alignement est obtenu avec NCBI Taxonomy au regard des correspondances directes et des distances d'édition des chaines alignées.

A noter qu'environ 9,4 % à 12,5% des concepts de la base AP-HP sont imparfaitement alignés avec NCBI Taxonomy ou la SNOMED CT : la distance de Levenshtein entre concept de la base AP-HP et proposition est supérieure à 20.

Ces résultats permettent de mettre en évidence les manques de SNOMED CT en termes de couverture et clarifient les règles éditoriales à mettre en œuvre pour améliorer les alignements en exact « match ».

[11] https://www.snomed.org/SNOMED/media/SNOMED/documents/IHTSDO-Affiliate-License-Agreement_UK_20200101-v1-0.pdf
[12] http://obofoundry.org/ontology/ncbitaxon

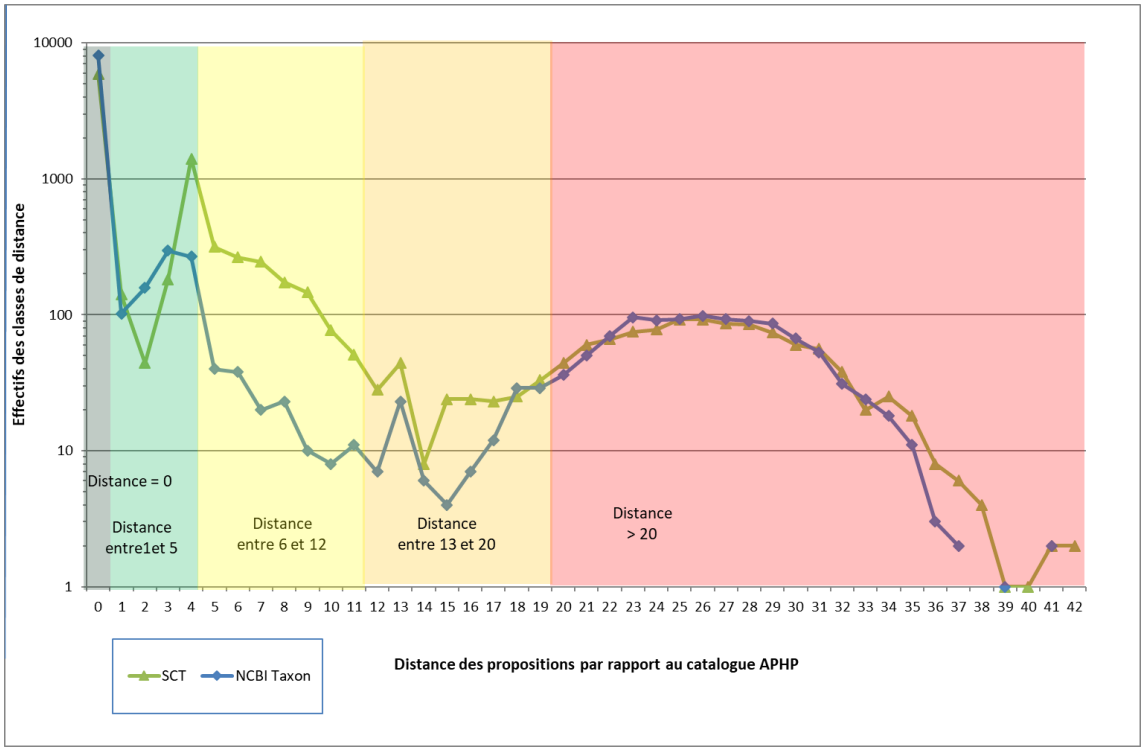
Tableau 6 : Comparaison des alignements de NCBI Taxonomy et SNOMED CT au catalogue de l'AP-HP (10146 bactéries) en fonction des distances de Levenshtein

	Alignement avec NCBI Taxonomy			Alignement avec SNOMED CT		
Distance de Levenshtein	**Distance moyenne**	**Nombre de concepts correspondants**	**% total**	**Distance moyenne**	**Nombre de concepts correspondants**	**% total**
A distance = 0	0	8040	**79,2%**	0	5900	**58,2%**
B distance entre 1 et 5	3	864	8,5%	4	2088	20,6%
C distance entre 6 et 12	8	117	1,2%	8	984	9,7%
D distance entre 13 et 20	17	146	1,4%	17	225	2,2%
E distance > 20	27	979	9,6%	27	949	9,4%
Distance moyenne (total base AP-HP)	3,2			4,4		

La figure ci-après illustre la distribution des effectifs dans chaque classe unitaire de distance éditoriale.

Les principales différences de courbe résident dans la première partie du graphe. Un plus grand nombre « d'exact matchs » pour NCBI Taxonomy a pour conséquence un plus grand nombre de propositions entre 3 et 20 pour SNOMED CT. Au-delà de 20 les courbes de SNOMED CT et NCBI Taxonomy sont superposables.

Figure 8 : Distribution des alignements de NCBI Taxonomy et SNOMED CT vs le catalogue AP-HP en fonction des distances de Levenshtein



4.2.4.2 Analyse qualitative

Les alignements qui requièrent au moins un ajustement (distance de Levenshtein > 0) dans la chaîne de caractère de référence (insertion, suppression ou changement) ont été analysés par zone de distance.

4.2.4.2.1 Zone de faible distance (1 à 5)

Respectivement 864 et 2088 alignements de la NCBI Taxonomy et SNOMED CT se trouvent dans cette zone de distance sémantique par rapport à la base AP-HP.

La principale différence entre SNOMED CT et NCBI Taxonomy réside dans le fait que SNOMED CT ne définit pas les espèces non dentifiées ou non encore décrites. Ceci est noté « sp. » pour espèce (ex : Alicycliphilus sp.). NCBI Taxonomy permet le codage à ce niveau, signifiant ainsi que l'espèce est en attente d'identification.

La base AP-HP répertorie 1531 espèces définies de cette manière. Elles ne peuvent être alignées avec la SNOMED CT qui propose alors un codage au niveau supérieur du genre (par exemple Alicycliphilus (organism), code : 426712005 pour aligner le concept AP-HP 1ALC_CLP Alicycliphilus sp.). Ceci entraîne une perte de précision de codage.

Les autres sources de mauvais alignement dans cette zone de distance proviennent de :
- Bactéries absentes de la terminologie (ex agromyces allii, dokdonia, yania flavia …) pour la SNOMED CT ;
- Syntaxe incluant des caractères spéciaux et des abréviations (ex : CDC group F-1) ;
- Emploi d'un synonyme au niveau du genre (ex : angiococcus/archangium) qui n'est plus employé au niveau de l'espèce (ex seul "Archangium sp." est répertorié dans NCBI Taxonomy).

Le tableau 7 illustre ces observations.

Tableau 7 : Exemples de problèmes d'alignement avec la base AP-HP pour des zones de distance faibles (entre 1 et 5)

Concept Base AP-HP	Alignement automatique NCBI Taxonomy	Distance au concept AP-HP	Alignement automatique SNOMED CT	Distance au concept AP-HP
Luteibacter sp.	luteibacter sp.	0	luteibacter *pas de possibilité de coder genre sp. dans SNOMED CT*	4
Agromyces allii	agromyces allii	0	~~agromyces albus~~	3
Yania flava	yania flava	0	~~tinea flava~~ *yania flava absent de SNOMED CT*	3
Dokdonia donghaensis	dokdonia donghaensis	0	~~gordonia aichiensis~~ *genre dokdonia absent de SNOMED CT*	5
Dokdonia sp.	dokdonia sp.	0	~~gordonia soli~~ *genre dokdonia absent de SNOMED CT*	5
CDC enteric group 63	cdc enteric group 63	0	~~cdc enteric group 69~~ *groupe 63 absent de SNOMED CT*	1
CDC enteric group 69	~~cdc enteric group 19~~ *groupe 69 absent de NCBI Taxonomy*	1	cdc enteric group 69	0
Angiococcus sp.	~~agrococcus sp.~~ *genre Angiococcus absent de NCBI Taxonomy remplacé par son synonyme Archangium*	2	angiococcus *pas de possibilité de coder genre sp. dans SNOMED CT*	4

NB : le tableau donne des exemples avec des distances de>0 dans une des 2 bases ; pour comparaison l'alignement dans l'autre base est donnée en exemple.
NB : le tableau donne des exemples avec des distances de>0 dans une des 2 bases ; pour comparaison l'alignement dans l'autre base est donnée en exemple.

4.2.4.2.2 Zone de distance intermédiaire (6 à 20)

Respectivement 203 et 1209 alignements de la NCBI Taxonomy et SNOMED CT se trouvent dans cette zone de distance sémantique par rapport à la base AP-HP.

Les problèmes d'alignements ont plusieurs causes :

- Des bactéries répertoriées par l'AP-HP absentes de la SNOMED CT (ex : conchiformibius kuhniae, Actinopycnidium caeruleum, Acidicaldus organivorans, Actinospica robiniae, Tamlana crocina…) ;
- Un emploi de synonymes par l'AP-HP au niveau du genre qui n'est pas employé pour l'espèce dans NCBI Taxonomy ou SNOMED CT (ex Ehrlichia/cowdria) ;
- Une ligne éditoriale de l'AP-HP qui a choisi d'apposer deux synonymes dans le même libellé, résultant en des ambigüités d'alignement. Dans ce cas les alignements sont souvent corrects mais restent dans des zones de distance élevées du fait de la suppression d'une partie de la chaîne de caractère.

Le tableau 8 illustre ces observations.

Tableau 8 : Exemple de problèmes d'alignement avec la base AP-HP pour des zones de distance intermédiaires (entre 5 et 20)

Concept Base AP-HP	Alignement automatique NCBI Taxonomy	Distance au concept AP-HP	Alignement automatique SNOMED CT	Distance au concept AP-HP
Acidicaldus organivorans	acidicaldus organivorans	0	~~acidiphilium organovorum~~ *acidicaldus organivorans absent de SNOMED CT*	9
Actinopycnidium caeruleum	actinopycnidium caeruleum	0	~~actinobacillus capsulatus~~ actinopycnidium caeruleum *absent de SNOMED CT*	11
Actinospica robiniae	actinospica robiniae	0	~~actinomyces radingae~~ actinospica robiniae *absent de SNOMED CT*	8
Conchiformibius kuhniae	conchiformibius kuhniae	0	~~moniliformis dubius~~ conchiformibius kuhniae *absent de SNOMED CT*	11
Tamlana crocina	tamlana crocina	0	~~porzana carolina~~ tamlana crocina *absent de SNOMED CT*	6
Cowdria sp.	~~chondria sp.~~ *Genre cowdria présent en tant que synonyme de du genre Ehrlichia dans NCBI Taxonomy. Synonyme non décliné pour nommer l'espèce*	2	~~Coenuriasis~~ *Genre ehrlichia présent dans SNOMED CT non identifié comme synonyme de cowdria*	6
Streptococcus agalactiae (groupe B)	Streptococcus agalactiae *Le synonyme S groupe b est absent de NCBI Taxonomy*	6	Streptococcus agalactiae *S groupe B est un synonyme.*	7
Gordonia (ou Gordona) rhizosphera	gordonia rhizosphera (*Correct)*	13	gordonia rhizosphera (correct)	13
Lactobacillus uli (nv : Olsenella uli)	Olsenella uli (*correct : nouveau nom)*	16	Lactobacillus uli (organism) (correct : ancien nom)	15
Enterobacter agglomerans (nv : Pantoea agglomerans)	enterobacter agglomerans *(correct : ancien nom)*	21	enterobacter agglomerans (correct : ancien nom)	18
Erwinia milletiae (nv : Enterobacter agglomerans)	enterobacter agglomerans *(correct nouveau nom mais ambiguité car doublon créé cf. ligne ci-dessus)*	24	~~erwinia herbicola-enterobacter agglomerans group~~ ambiguité a mené à un alignement vers un groupe de bactéries	20

NB : le tableau donne des exemples avec des distances de 6 à 20 dans une des 2 bases ; pour comparaison l'alignement dans l'autre base est donnée en exemple.

4.2.4.2.3 Zone des grandes distances (>20)

Respectivement 979 et 949 alignements de NCBI Taxonomy et SNOMED CT se trouvent dans cette zone de distance sémantique par rapport à la base AP-HP.

Tous les alignements rencontrent la problématique éditoriale du choix d'un ancien nom vs le nouveau. Ceci limite considérablement les possibilités d'alignement automatique et de son contrôle par les distances.

Le tableau 9 illustre quelques exemples.

Tableau 9 : Exemples de problèmes d'alignements avec la base AP-HP pour des zones de grandes distances (>20)

Concept Base AP-HP	Alignement automatique NCBI Taxonomy	Distance au concept AP-HP	Alignement automatique SNOMED CT	Distance au concept AP-HP
Alcaligenes xylosoxidans/denitrificans (nv : Achromobacter)	alcaligenes xylosoxydans denitrificans *(correct : ancien nom)*	23	alcaligenes xylosoxidans subsp. Denitrificans *(correct : ancien nom)*	26
Bacteroides ruminicola/brevis (nv : Prevotella ruminicola/brevis)	bacteroides ruminicola subsp. Brevis *(correct : ancien nom)*	35	Bacteroides ruminicola ss. Brevis *(correct : ancien nom)*	37
CDC enteric group 64 (nv : Buttiauxiella gaviniae)	buttiauxella gaviniae *(correct : nouveau nom)*	28	buttiauxella gaviniae *(correct : nouveau nom)*	28
Mycoplasma lactucae (nv : Mesoplasma lactucae)	mesoplasma lactucae *(correct : nouveau nom)*	23	mesoplasma lactucae *(correct : nouveau nom)*	22

5 DISCUSSION

5.1 Terminologies de microbiologie

Cette étude a montré que de nombreuses ressources sémantiques contiennent des microorganismes.

En effet, la recherche systématique d'un échantillon de 15 bactéries issues du catalogue de germes de l'APHP Sur le serveur multi terminologies BIOPORTAL a permis d'identifier 24 ressources (parmi les 845 contenue dans BIOPORTAL) contenant tout ou partie de l'échantillon.

6 sont intéressantes pour répertorier des germes Communs : NCBI Taxonomy, SNOMED CT, LOINC, NCIT, CIM-11 et MESH.

Par rapport aux besoins de l'AP-HP (couverture d'un catalogue), 2 ont mérité une investigation approfondie : NCBI Taxonomy, référence mondiale spécialisée dans le référencement du vivant et SNOMED CT, terminologie multi-domaines adoptée dans 39 pays.

5.2 Audit de NCBI Taxonomy et SNOMED CT

Comme dans d'autres études de ce rapport (annexe P4.8 Médicament, annexe P4.2 Anatomie, annexe P4.3 SLA (maladie de Charcot)), la SNOMED CT montre ses limites en termes de couverture par rapport à des terminologies spécialisées. **La couverture de la SNOMED CT dans le domaine des bactéries est ainsi quantitativement largement inférieure à celle de NCBI Taxonomy : 11 508 termes contre 371 766.**

Du point de vue opérationnel, La SNOMED CT contient 6391 bactéries nommées au niveau genre espèce. En effet, il faut éliminer des termes pré-coordonnés qui définissent des nuances de sérogroupes mais qui apportent de la redondance taxonomique et qui sont trop spécifiques en annotation. Cet univers est insuffisant par rapport au catalogue de l'AP-HP qui contient 10 146 bactéries.

NCBI Taxonomy contient 91 442 bactéries formellement identifiées au niveau genre espèce (14 fois plus que SNOMED CT). A noter qu'un bruit de fond de 280 324 bactéries est apporté par des souches en cours d'identification, notée « sp. » suivi d'une séquence alphanumérique.

SNOMED CT et NCBI Taxonomy ont des règles de présentation des bactéries différentes :

- Plus grande richesse de synonymes dans NCBI Taxonomy (cf. Nesseiria Meningitidis, Bacteroides Fragili) ;
- Bactéries présentées sous leur nom usuel dans SNOMED CT (ex : bacille du charbon, méningocoque).

Ceci montre que NCBI Taxonomy est plus précise que SNOMED CT pour structurer des résultats microbiologiques dans les règles de l'art, mais aussi préfigure la complémentarité que peuvent avoir ces deux terminologies dans le cas d'usage « annotation » sur des dossiers dont nous ignorons comment sont structurées les informations microbiologiques.

Ces différences de lignes éditoriales seraient à approfondir pour mieux positionner ces 2 terminologies l'une par rapport à l'autre dans le cas d'usage annotation.

Du point de vue Structure, NCBI Taxonomy est un arbre taxonomique monoaxial adapté pour classer les bactéries.

L'arbre de la SNOMED CT mixe taxonomie, caractéristiques morphologiques et propriétés biochimiques, apportant une plus grande richesse de description.

Cette richesse peut également être atteinte en mêlant NCBI Taxonomy avec des terminologies apportant ces précisions : par exemple LOINC, NCIT, Ontology of microbial Phenotypes[13] ou microbial phenotype ontology[14].

[13] https://bioportal.bioontology.org/ontologies/OMP?p=summary
[14] https://bioportal.bioontology.org/ontologies/OMP?p=summary

Toutefois la richesse de classification de la SNOMED CT au niveau des bactéries n'est pas a priori uniforme sur tout le domaine microbiologique. Ceci est lié au fait que la SNOMED CT n'est pas une ontologie formelle. Cette observation limite l'utilisation de la SNOMED CT sur des cas d'usage de « raisonnement automatique ». Une analyse approfondie est nécessaire pour définir les limites d'utilisation de la SNOMED CT dans ce cas d'usage et éventuellement pour élaborer des solutions.

Par rapport aux standards de qualité du web sémantique (Web des données), les données de NCBI Taxonomy présentent l'intérêt d'être liées. NCBI Taxonomy est un « *hub* » de connaissance ouvert vers d'autres ressources. Les identifiants URIs NCBI Taxonomy permettent ainsi de mettre en contexte les simples résultats d'une analyse microbiologique. A noter que l'identifiant taxon (Taxid :) est un identifiant reconnu par d'autre bases biologiques (ex : Backdive ABacterial diversity metadatabase). A l'opposé, SNOMED CT, placée dans un modèle propriétaire, ne présente pas de passerelle de connaissances vers d'autres bases. Ceci est principalement dû au fait que le SCTID (identifiant primaire de la SNOMED CT) ne peut être réutilisé sans licence.

Les données liées sont importantes pour les communautés scientifiques et de l'interopérabilité. Il s'agit d'un ensemble de principes de conception pour le partage de données sur le web. La modélisation des données en RDF, leur recherche en SPARQL et leur localisation par URI sont la fondation des données liées.

Les données liées sont des catalyseurs pour l'interopérabilité, permettant l'intégration de données avec peu d'impact sur les systèmes existants. Elles permettent créativité et innovation dans le contexte de création de connaissance.

Le tableau ci-après synthétise les principales observations.

Tableau 10 : Comparaison NCBI Taxonomy et SNOMED CT en termes de couverture et d'exploitabilité

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 pré-coordonnés 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
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
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

5.3 ALIGNEMENTS VS CATALOGUE DE L'APHP

Le meilleur alignement avec le catalogue de l'AP-HP est obtenu avec NCBI Taxonomy au regard des correspondances directes.

Au total, respectivement 79.2% et 58.2% du catalogue AP-HP sont directement alignés avec NCBI Taxonomy et SNOMED CT (étiquettes principales ou synonymes).

A noter que l'alignement avec SNOMED CT ne pourra pas être grandement amélioré en raison du nombre limité de bactéries structurées au niveau genre-espèce (6391).

Ces résultats permettent de mettre en évidence les manques de SNOMED CT en termes de couverture et clarifient les règles éditoriales à mettre en œuvre pour améliorer les alignements en « exact-matchs ».

Tableau 11: Evaluation de l'alignement du catalogue AP-HP avec NCBI Taxonomy et SNOMED

Critère	SNOMED CT	NCBI Taxonomy	Conclusion
► Alignement de la terminologie (par rapport au catalogue AP-HP)	**58.2% Exact match (5900 termes)**	**79.2% Exact match (8040 termes)**	► NCBI Taxonomy présente une meilleure qualité d'alignement que SNOMED CT

5.4 Accessibilité et exploitabilité

Le principal inconvénient de SNOMED CT est son format propriétaire qui limite l'accessibilité et l'exploitabilité des données par rapport à l'offre de terminologies en open data, notamment en cas de résiliation de licence[15] (Voir aussi l'étude juridique menée dans le cadre de ce rapport : annexe P2.0).

Du point de vue accessibilité et exploitabilité, NCBI Taxonomy est nettement supérieure à SNOMED CT qui dans le cas de la microbiologie, est un catalogue privé exploitant des données publiques.

A noter qu'il existe de nombreuses initiatives publiques, dont la fonderie d'Ontologies Biomédicales ouverte (OBO Foundry) proposant des solutions ouvertes pour distribuer des ressources sémantiques.

OBO Foundry (http://obofoundry.org/) intègre NCBI Taxonomy et propose une approche où les modèles de conception et les meilleures pratiques de spécification de ressources sémantiques sont promus sur une base de développement collaboratif et de libre accessibilité.

OBO Foundry a établi un ensemble de principes fédérateurs[16] pour les ontologies intégrant le projet. Notamment il est stipulé que l'ontologie soit ouverte et disponible sans autre contrainte que de reconnaître la paternité à l'unité de production sémantique.

Ces principes sont également porteurs de normes de qualité, de rigueur formelle dans la représentation des données, ainsi que d'interopérabilité entre les ontologies ayant intégré l'OBO Foundry.

[15] https://www.snomed.org/SNOMED/media/SNOMED/documents/IHTSDO-Affiliate-License-Agreement_UK_20200101-v1-0.pdf (clause 5.6)
[16] http://obofoundry.org/principles/fp-000-summary.html

Tableau 12 : Evaluation de l'accessibilité et de l'exploitabilité des vocabulaires

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

5.5 Limites et perspectives de l'étude

L'évaluation de la couverture des terminologies de référence candidates a été réalisée de manière automatique.

⇨ Les alignements automatiques avec NCBI Taxonomy vont être revus et finalisés par des experts en microbiologie assistés par des biologistes experts en interopérabilité permettant la standardisation du catalogue de souches bactériennes de l'AP-HP.

L'évaluation de la couverture des terminologies de référence candidates a été réalisée dans le seul contexte de l'AP-HP (spécifique ? généralisation de l'analyse ?).

⇨ Extension de l'évaluation de la terminologie NCBI Taxonomy en conditions d'utilisation réelle en collaboration avec les partenaires hospitaliers dans le cadre du réseau ONERBA. Les efforts vont porter en priorité sur les analyses fréquentes.

Le périmètre du présent travail n'a pas adressé la standardisation des résultats d'antibiogramme. Ce sujet va être traité.

Par ailleurs, un autre cas d'usage des résultats d'analyses microbiologiques va être considéré. Il s'agit d'analyser la pertinence de l'usage de la terminologie NCBI Taxonomy dans le cadre de l'aide au codage PMSI c'est-à-dire lorsqu'il s'agit d'identifier automatiquement des diagnostics CIM-10 d'antibiorésistance à partir de résultats d'analyses microbiologiques.

L'utilisation conjointe avec la LOINC doit être testée.

6 CONCLUSION

Par rapport à l'objectif de rechercher une terminologie de référence pour coder le catalogue de germes de l'AP-HP, cette étude a permis d'identifier 24 terminologies contenant des bactéries.

Parmi ces ressources NCBI Taxonomy est la meilleure pour répondre aux besoins de l'AP-HP de constituer un vocabulaire commun de description des microorganismes.

Elle se distingue de SNOMED CT en terme :

- De meilleure couverture du domaine bactérien ;
- De Meilleure précision de nommage des bactéries ;
- De Meilleure correspondance directe avec le catalogue de l'AP-HP ;
- De Meilleure ouverture des données (Modèle ouvert vs Modèle propriétaire pour SNOMED CT) ;
- De données pouvant être ouvertes et liées vers d'autres base de connaissances.

Cette étude a également montré que NCBI Taxonomy et SNOMED CT pouvaient être complémentaires sur des cas d'usage d'annotation pour lesquels une approche multi-terminologique donne les meilleurs résultats.

7 ANNEXE : DESCRIPTIONS DES TERMINOLOGIES CONTENANT DES BACTERIES

L'échantillon de 15 bactéries a été retrouvé quasi intégralement dans deux terminologies : SNOMED CT et NCBI Taxonomy (cf. supra).

A noter qu'une partie des bactéries a été retrouvée dans 22 autres terminologies qui sont décrites dans cette annexe.

Un focus est fait sur les 4 terminologies répertoriant entre 10 et 11 des 15 bactéries issues de l'échantillon AP-HP (NCIT, CIM-11, LOINC et MeSH). Les autres terminologies sont brièvement présentées dans le tableau 13 en fin d'annexe.

7.1 National Cancer Institute Thesaurus (NCIT)[xvii]

Le NCIT est une ressource de référence qui couvre le domaine du cancer dans de multiples dimensions (pathologie, signes, anatomie ; gènes, médicaments, agents causaux, actes, etc…). Le NCIT s'appuie sur des terminologies existantes en établissant des passerelles. Le thésaurus sert ainsi de hub de connaissance grâce à ses relations sémantiques.

Le Thésaurus contient actuellement plus de 154 947 concepts, structurés en 18 sections (source Bioportal). Plus de 115 000 concepts sont associés à des définitions textuelles. Le réseau sémantique est riche de 400 000 relations inter concepts.

Les bactéries sont regroupées dans le chapitre organisme.

Les bactéries répertoriées dans le NCIT sont complètement alignées avec NCBI Taxonomy.

Le NCI Thesaurus est produit par l'équipe « Enterprise Vocabulary Services » (EVS)[xviii] du National Cancer Institute, (Ma. USA).

Le thésaurus NCI est publié sous licence ouverte Creative Commons Attribution 4.0 (CC-BY v4.0).

7.2 Logical Observation Identifiers Names & Codes (LOINC)

La LOINC (Logical Observation Identifiers Names and Codes) est une terminologie internationale de référence pour le codage des observations médicales et des documents électroniques, publiée par le Regenstrief Institute (organisation de recherche médicale à but non lucratif).

Elle é été créée en 1994 pour fournir une base ouverte codant des soins cliniques et des résultats de laboratoire.

La LOINC contient 87 863 concepts codés (v 2.64) répartis en 362 classes. La classe microbiologie (MICRO) regroupe 12 251 termes. Elle est la plus importante classe au sein de la LOINC.

La LOINC Liste environ 580 bactéries intégrées dans 3250 codes LOINC (demande d'examen et observations).

La licence de la LOINC est une licence spécifique du Regenstrief Institute[xix] permettant le libre accès à la LOINC sans modification de son contenu.

7.3 CIM-11

La CIM-11[xx] est la nouvelle version de la Classification internationale des maladies produite par l'OMS. La première version stable de la CIM-11 a été publiée en 2018.

Elle est définie sur un modèle ontologique. Elle est riche de 56 000 concepts formellement définis complétés par plus de 70 000 termes indexés (synonymes et étiquettes alternatives).

La CIM-11 regroupe 791 agents infectieux (bactéries, virus, parasites, champignons, prions…) dans son chapitre « Extension Codes ».

Les concepts de microorganismes sont tournés vers la post coordination : combinaison de codes extensions avec des concepts de situations cliniques pour en préciser l'étiologie. Par exemple le concept de Klebsiella Pneumoniae (code XN741 dans la linéarisation MMS) précisera l'origine bactérienne d'une pneumonie (code CA40.0Y dans la linéarisation MMS).

334 concepts codés sont référencés pour décrire les bactéries avec 4 échelons dans l'arborescence :

- Le niveau 1 définit les caractéristiques de la paroi bactérienne (gram positif, gram négatif ou ni gram négatif, ni gram positif) ;
- Le niveau 2 définit directement le genre (ex : yersinia, streptococcus, etc.)[17];
- Les niveaux 3 et 4 définissent les espèces et sous espèces (ex : Streptococcus Mutans).

Une liste de 273 bactéries définies au niveau espèce ou sous espèces est ainsi rendue disponible.

Ces bactéries sont définies par une étiquette simple sans synonyme ou étiquette alternative : le genre streptomyces sera appelé par ce seul nom et n'aura pas la correspondance avec Chainia qui est synonyme dans d'autres terminologies.

La politique de publication de la CIM-11 est en cours d'élaboration entre l'OMS, les états membres et les centres collaborateurs (INSERM-CepiDc pour la France). La diffusion de la CIM-11 se fera sous licence CC-BY-ND[18].

[17] Exception faite de Kingella kingae décrite au 2eme niveau. Il s'agit d'une bactérie de la flore oropharyngée pouvant être responsable d'infection ostéo-articulaires chez l'enfant.
[18] https://icd11files.blob.core.windows.net/refguide/html/index.html#applicability-and-intellectual-property

7.4 Thésaurus MeSH[xxi]

Le MeSH (Medical Subject Headings) est le thésaurus de référence dans le domaine biomédical. Il s'agit d'un vocabulaire de mots-clés qui permettent de décrire le contenu des articles de la base Medline. Il est traduit en français par l'INSERM.

Les termes MeSH (descripteurs) sont répartis en 16 grandes catégories, structurées en arborescence (des termes plus génériques aux termes plus spécifiques).
Les bactéries sont une sous-catégorie de la catégorie organisme. Elles sont elles-mêmes divisées en 19 sections (http://mesh.inserm.fr/FrenchMeSH/view/index.jsp).

Le cas d'usage principal du MeSH est la description de la totalité des articles référencés dans la base Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/). Le vocabulaire MeSH permet de construire des requêtes cohérentes et précises. Le MeSH se développe en continu pour indexer les articles les plus récents.

Le MeSH français est disponible sous forme de fichier xml (http://mesh.inserm.fr/FrenchMesh/conditions_en.htm) sous licence libre Creative Common avec attribution de paternité et usage non commercial (CC-BY-NC).

7.5 Autres terminologies

Le tableau 13 ci-après décrit les 18 autres terminologies dans lesquelles sont répertoriées des bactéries de l'échantillon.

Ces terminologies ont des objectifs divers :

- Indexer des documents ou des travaux de recherche comme le fait le dictionnaire MeSH : CRISP, PLOSTHES et SYN. Ces 3 terminologies sont issues de projets collaboratifs ;
- Décrire des domaines de connaissance spécifiques : AHOL, CSEO, Fly Taxonomy, ONTOLurgences, Ontopneumo, OntoAD, RGO ;
- Servir de passerelle entre différents domaines de connaissance pour les lier ou les intégrer entre eux : IOBC, OCHV, ORTH ;
- Être des terminologies de référence dans leur domaine : CPT (actes), NDDF et RxNORM (médicament), CTV3 et SNMI (multi-domaines).

La majorité de ces terminologies est accessible sous licence ouverte (AHOL, CSEO, CTV3, PLOSTHES et SYN) ou de libre diffusion (CRISP, IOBC, ONTOLurgences, Onto Pneumo, RGO, RxNORM.

Trois sont distribuées sous licence propriétaire. Toutes sont des terminologies américaines. NDDF est une terminologie du médicament comparable aux bases de médicaments commerciales françaises de type VIDAL ou BCB. SNMI est la version antérieure de SNOMED CT. CPT est comparable aux terminologies d'acte (CCAM ou ICHI) dans son contenu. Elle est produite et commercialisée par l'association médicale américaine (AMA).

L'information sur les licences n'est pas disponible pour trois d'entre elles : Fly Taxonomy, OCHV et OntoAD.

Tableau 133 : Terminologie incluant des bactéries de l'échantillon : Description et publication

Terminologie Unité de production Volumétrie Téléchargement de la terminologie	Description	Domaines bactéries et microorganismes	Licence
AHOL - Animal Health Ontology for Livestock LOVINRA (FR) 339 classes codées Accessible sur Bioportal	AHOL est une ontologie des caractéristiques définissant les problèmes de santé des animaux d'élevage dans leur environnement (EOL) liés à leurs phénotypes (ATOL)[xxii]. LOVInra recense les vocabulaires (ontologies, thésaurus, terminologies, etc.) produits à l'INRA et les publie selon les principes des Linked Open Data.	Les bactéries sont classées au niveau genre espèce dans le domaine organisme. **Les bactéries présentes dans cette terminologie sont complètement alignées avec NCBI Taxonomy.**	Licence ouverte
CRISP - Computer Retrieval of Information on Scientific Projects NIH (USA) 9039 classes codées Accessible sur Bioportal	CRISP a été développé par le NIH[xxiii] pour indexer l'information biomédicale contenue dans une base de projets soutenus par les services de santé publique des USA (PHS) CRISP est organisé autour de 11 domaines.	Les bactéries sont classées au niveau genre espèce dans le domaine organisme.	Licence UMLS
CSEO - Cigarette Smoke Exposure Ontology 20 085 classes codées Accessible sur Bioportal	CSEO est une ontologie[xxiv] spécialisée de l'exposition à la fumée de cigarette avec un focus sur la description des éléments expérimentaux et sur l'impact de cette exposition environnementale	Les bactéries sont regroupées dans un domaine organisme.	Domaine public (pas de licence identifiée)[xxv]
CPT - Current Procedural Terminology American Medical association (AMA) (USA) 14 183 classes codées. Non accessible	CPT [xxvi]est une terminologie permettant le codage des actes de médecine de chirurgie et de diagnostics. Les codes CPT sont utilisés dans des cas d'usage administratifs par les assurances : traitement des demandes de remboursement et ententes préalables. CPT est la terminologie de référence, pour les actes aux USA.	Les bactéries sont associées à des procédures : par exemple leur identification ou les caractérisations de leurs propriétés.	Licence propriétaire
CTV3- Read Codes, Clinical Terms Version 3 NHS Digital (UK) 140 000 classes codées Accessible sur NHS Didital par le service TRUD (Technology reference data update distribution)	Les read codes sont un thésaurus codé de termes cliniques. Ils sont utilisés dans le NHS depuis 1985. Ils fournissent un vocabulaire standard permettant aux cliniciens de structurer leur pratique (observations et actes) [xxvii]. Les Read codes sont désormais intégrés à la SNOMED CT. Ils peuvent encore être utilisés en dentisterie et psychiatrie jusqu'en 2020[xxviii]	Les bactéries sont classées selon leurs formes et propriétés.	Open government licence
Fly Taxonomy – 6 600 classes codées G Baechli et al (U Zurich) Accessible sur Bioportal	Taxonomie de la famille des Drosophilidae (en grande partie d'après Baechli et al.et d'autres taxons mentionnés dans FlyBase. Cette ontologie embarque des micro-organismes dont des bactéries.	Les bactéries présentes sont classées en genre et espèces. Elles sont **complètement alignées avec NCBI Taxonomy.**	Pas d'information
IOBC - Interlinking Ontology for Biological Concepts National Bioscience Database Center (Japon) 126 842 classes codées Accessible sur Bioportal	IOBC est une terminologie multi domaine contenant des concepts biologiques. La terminologie est disponible en anglais.[xxix]	Les bactéries sont classées le plus souvent au niveau du genre, avec peu de profondeur hiérarchique.	CC-BY-NC (usages non commerciaux)
NDDF – National drug data file First data bank (USA) 29 887 classes codées Non accessible	Base de connaissance du médicament (indications, CI, Effets indésirables, interactions…) Terminologie obsolète remplacée par FDBMedknowledge[xxx] non référencée sur Bioportal.	Pas d'information	Propriétaire
OCHV - Ontology of Consumer Health Vocabulary University of Utah (USA) 115 645 classes codées Accessible sur Bioportal	Open-access collaborative CHV est une ontologie contenant des termes employés par des consommateurs de soins. Elle est obtenue par extraction de concepts récurrents sur des sites grand public. Elle est destinée à s'interfacer avec du langage technique pour en faciliter la traduction en langage profane.	Les bactéries sont présentées en listes simples sans hiérarchies.	Pas d'information

Terminologie Unité de production Volumétrie Téléchargement de la terminologie	Description	Domaines bactéries et microorganismes	Licence
ONTOLUrgences J Charlet _UPMC Univ Paris 06, UMR_S 1142, LIMICS, F-75006 (FR) 10 031 classes codées Accessible sur Bioportal	ONTOLUrgences [xxxi] est une ontologie développée pour indexer et récupérer des informations dans le dossier médical d'urgence électronique d'un patient.	Les bactéries sont classées au niveau du genre ou genre et espèce, ou genre espèce avec caractéristique de résistance (adaptée au cas d'usage soin en urgence). L'échantillon de bactéries est très réduit (3 bactéries)	CC-BY-NC-ND[xxxii] (usages non commerciaux, pas d'œuvre dérivée)
Ontopneumo A Baneyx _INSERM, UMR_S 872, LIMICS, F-75006 (FR) 1 153 classes codées Accessible sur Bioportal	OntoPneumo est une ontologie pour décrire la pneumologie (version française L'ontologie a été produite par traitement du langage naturel sur des documents médicaux. OntoPneumo est utilisé dans un logiciel pour pneumologues pour supporter le traitement des séjours.	Ontopneumo liste les bactéries et autres micor-organismes interagissant avec l'appareil respiratoire	CC-BY-NC-ND (usages non commerciaux, pas d'œuvre dérivée)
OntoAD Bilingual Ontology of Alzheimer's Disease and Related Diseases Khadim Dramé ISPED Bordeaux II (FR) 5 899 classes codées Accessible sur Bioportal	OntoAD[xxxiii] est une ontologie de domaine bilingue (anglais-français) pour modéliser les connaissances sur la maladie d'Alzheimer et les syndromes associés.	Les bactéries sont regroupées dans la section dédiée aux résultats de laboratoire	Pas d'information
ORTH – Orthology Ontology Initiative internationale 496 4 classes codées Accessible sur Bioportal	ORTH est destinée [xxxiv, xxxv] à normaliser sémantiquement le domaine orthologique (liens évolutifs entre gènes). De création récente (2015) elle contribuera à une meilleure exploitation des ressources orthologiques dans la recherche biomédicale.	Les bactéries sont incluses au fur et à mesures que des liens orthologiques sont mis en évidence. **Complètement alignées sur NCBI Taxonomy.**	Licence ouverte
PLOSTHES - PLOS Thesaurus Public Library of Science (USA/UK) 10 712 classes codées Accessible sur Bioportal	PLOS est un éditeur engagé dans l'open access en tant que vecteur d'innovation[xxxvi]. Les termes de PLOSTHES couvrent l'éventail de sujets de recherche inclues dans les revues de PLOS. Il a été initialement construit en 2012 pour indexer des articles de recherche.	PLOSTHES contient un échantillon de bactéries communes codées au niveau du genre ou de l'espèce. Il n'y a pas de hiérarchie organisée.	CC-BY
RGO - Radiology Gamuts Ontology C Kahn –Univ Pennsylvannie (USA) 18 001 classes codées Accessible sur Bioportal	RGO est une ressource pour catégoriser et lier des diagnostics radiologiques. RGO contient plus de 2 000 listes de diagnostics différentiels pour les résultats d'imagerie dans tous les systèmes corporels. Affichage des causes pour une constatation puis affichage de toutes les constatations associées à une cause. Vous pouvez afficher toutes les causes d'une constatation, puis cliquer sur l'un des diagnostics pour afficher toutes les constatations qu'elle provoque.	Certaines étiologies bactériennes font parties des causes associées à une constatation	CC-BY-NC-ND (usages non commerciaux, pas d'œuvre dérivée)
RxNORM – Terminologie de médicament (USA) National Library of medecine 113 182 classes codées Accessible sur Bioportal	RxNORM contient tous les médicaments disponibles aux USA. Elle fait partie des terminologies de l'UMLS[xxxvii]. RxNORM établit des passerelles entre différentes terminologies de médicaments et a donc un rôle en interopérabilité.	Les bactéries présentes dans RxNORM (114) sont référencées en tant qu'ingrédients (Termtype IN).	Licence UMLS gratuite

Terminologie Unité de production Volumétrie Téléchargement de la terminologie	Description	Domaines bactéries et microorganismes	Licence
SNMI - Systematized Nomenclature of Medicine, International Version SNOMED International (USA/UK) 109 150 classes codées Accessible sur Bioportal	Version antérieure de la terminologie SNOMED. Remplacée par la SNOMED CT Version dépréciée en 2017[xxxviii]	Les bactéries sont classées dans la section organismes vivants. Elles sont principalement codées au niveau de genre.	Propriétaire
SYN – Sage Bionetworks Synapse Ontology 14 462 classes codées Accessible sur Bioportal	Synapse [xxxix]est une plateforme de recherche collaborative qui permet aux équipes de partager, suivre et discuter de leurs données et analyses dans des projets. Synapse est aussi un web service avec un accès via une API Rest. L'ontologie permet d'annoter des données des travaux.	Les bactéries présentes sont regroupées dans la section organismes et codées dans une arborescence taxonomique	CC-BY

8 BIBLIOGRAPHIE

[1]European Centre for Disease Prevention and Control
European Antimicrobial Resistance Surveillance Network (EARS-Net)
https://www.ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/ears-net

[2]David Trystram, Emmanuelle Varon, Yves Péan, Hajo Grundmann, Laurent Gutmann, Vincent Jarlier, Hélène, Aubry-Damon.
Réseau européen de surveillance de la résistance bactérienne aux antibiotiques (EARSS) : résultats 2002, place de la France
BEH n° 32-33/2004, p141-144

[3]Robert J1, Veziris N; réseau Azay-Mycobactérie et le Conseil Scientifique de l'ONERBA.
Resistance to antituberculosis drug in France; data provided by the Azay-Mycobacteria network and the National Reference Center on Mycobacteria].
Med Mal Infect. 2008 Jun;38 Suppl 2:S68-70.

[4]Centre de ressource biologique de l'institut Pasteur (CRBIP)
Collection de l'institut Pasteur
https://research.pasteur.fr/fr/team/crbip/

[5]Macary F, DRON JC et al
Guide de dématérialisation des demandes et des résultats de bactériologie et parasitologie-mycologie
Guide_Bactériomycoparasito_20141003_V0039.Docx,
http://www.interopsante.org/form/412/1504/questionnaire-livre-blancguide-mycobacterioparasitologie.html

[6]Scott Federhen
The NCBI Taxonomy database
Nucleic acids Research, 2012, 40, D140-D143

[7]Whetzel PLet al
BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications
Nucleic Acids Res. 2011 Jul 1; 39(Web Server issue): W541–W545.

[8]Barros M, Couto FM.
Knowledge Representation and Management: a Linked Data Perspective
Yearb Med Inform, 2016, Nov 10;(1):178-183.

[9]Levenshtein,
Binary Codes Capable of Correcting Deletions, Insertions and Reversals,
Soviet Physics Doklady. 1966, 10, 707–710.

[10]Kiourtis A, Nifakos S, Mavrogiorgou A, Kyriazis D.
Aggregating the syntactic and semantic similarity of healthcare data towards their transformation to HL7 FHIR through ontology matching.
Int J Med Inform. 2019 Dec; 132:104002.

[11]Yagahara A, Uesugi M, Yokoi H.
Evaluation of Similar Term Definitions in Medical Device Adverse Event Terminology.
Stud Health Technol Inform. 2019 Aug 21; 264:1620-1621

[12]Ho T, Oh SR, Kim H.
A parallel approximate string matching under Levenshtein distance on graphics processing units using warp-shuffle operations.
PLoS One, 2017, Oct 10, 12(10).

[13]Workman TE1, Shao Y2, Divita G3, Zeng-Treitler Q2.
An efficient prototype method to identify and correct misspellings in clinical text.

BMC Res Notes. 2019, Jan 18,12(1), 42.

[xiv]The International Health Terminology Standards Development Organisation.
SNOMED Clinical Terms User Guide; 2017.
https://confluence.ihtsdotools.org/display/DOCSTART/SNOMED+CT+Starter+Guide?preview=/%2028742871/47677485/doc_StarterGuide_Current-en-US_INT_20170728.pdf

[xv]Gupta RS
The phylogeny and signature sequences characteristics of Fibrobacteres, Chlorobi, and Bacteroidetes.
Crit Rev Microbiol. 2004;30(2):123-43. Review.

[xvi]Wagner M, Horn M.
The Planctomycetes, Verrucomicrobia, Chlamydiae and sister phyla comprise a superphylum with biotechnological and medical relevance.
Curr Opin Biotechnol. 2006 Jun;17(3):241-9. Epub 2006 May 15. Review.

[xvii] https://ncit.nci.nih.gov/ncitbrowser/
[xviii] https://evs.nci.nih.gov/
[xix] https://loinc.org/license/
[xx] https://icd.who.int/dev11/l-m/en
[xxi] https://www.ncbi.nlm.nih.gov/mesh
[xxii] https://lovinra.inrae.fr/2019/12/20/animal-health-ontology-for-livestock/
[xxiii] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CSP/index.html
[xxiv] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120729/#!po=2.94118
[xxv] https://publicwiki-01.fraunhofer.de/CSEO-Wiki/index.php/CSEO_access
[xxvi] https://www.ama-assn.org/practice-management/cpt/cpt-overview-and-code-approval
[xxvii] https://digital.nhs.uk/services/terminology-and-classifications/read-codes
[xxviii] https://data.gov.uk/dataset/f262aa32-9c4e-44f1-99eb-4900deada7a4/uk-read-code
[xxix] https://bioportal.bioontology.org/ontologies/IOBC?p=summary
[xxx] https://www.fdbhealth.com/solutions/medknowledge
[xxxi] http://ebooks.iospress.nl/publication/37645
[xxxii] http://pertomed.limics.upmc.fr/~onto/ontologies/public/OntolUrgences/licenceOntolUrgences_2014-06-20.pdf
[xxxiii] https://bioportal.bioontology.org/ontologies/ONTOAD/?p=summary
[xxxiv] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4893294/
[xxxv] http://qfo.github.io/OrthologyOntology/
[xxxvi] https://www.plos.org/who-we-are
[xxxvii] https://www.nlm.nih.gov/research/umls/rxnorm/index.html
[xxxviii] https://en.wikipedia.org/wiki/Systematized_Nomenclature_of_Medicine
[xxxix] https://sagebionetworks.org/