Intégration de connaissances biomédicales hétérogènes grâce à un modèle basé sur les ontologies de support

Jean Nikiema

École doctorale
Sociétés, politique,
santé publique

Université de Bordeaux

École doctorale **Sociétés, politique, santé publique**
Unité de recherche **Bordeaux Population Health Research Center (BPH)**

Thèse présentée par **Jean Noël Nikiema**
Soutenue le **10 octobre 2019**
En vue de l'obtention du grade de docteur de l'Université de Bordeaux

Discipline **Santé publique**
Spécialité **Informatique et santé**

Titre de la thèse

Intégration de connaissances biomédicales hétérogènes grâce à un modèle basé sur les ontologies de support

Thèse dirigée par	Fleur Mougin	directeur
	Vianney Jouhet	co-directeur

Composition du jury

Rapporteurs	Anita Burgun	professeure au Paris Descartes University France	
	Stefan Schulz	professeur au Medical University Graz Austria	
Examinateurs	Olivier Bodenreider	directeur de recherche au NLM/NIH USA	
	Geneviève Chene	professeure au Bordeaux University France	présidente du jury
Directeurs de thèse	Fleur Mougin	MCU, HDR au Bordeaux University France	
	Vianney Jouhet	PH au CHU of Bordeaux	

École doctorale
Sociétés, politique, santé publique

université de BORDEAUX

Université de Bordeaux

École doctorale **Sociétés, politique, santé publique**
Unité de recherche **Bordeaux Population Health Research Center (BPH)**

Thèse présentée par **Jean Noël Nikiema**
Soutenue le **10 octobre 2019**
En vue de l'obtention du grade de docteur de l'Université de Bordeaux

Discipline **Santé publique**
Spécialité **Informatique et santé**

Titre de la thèse

Intégration de connaissances biomédicales hétérogènes grâce à un modèle basé sur les ontologies de support

Thèse dirigée par	Fleur Mougin	directeur
	Vianney Jouhet	co-directeur

Composition du jury

Rapporteurs	Anita Burgun	professeure au Paris Descartes University France	
	Stefan Schulz	professeur au Medical University Graz Austria	
Examinateurs	Olivier Bodenreider	directeur de recherche au NLM/NIH USA	
	Geneviève Chene	professeure au Bordeaux University France	présidente du jury
Directeurs de thèse	Fleur Mougin	MCU, HDR au Bordeaux University France	
	Vianney Jouhet	PH au CHU of Bordeaux	

Université de Bordeaux

Doctoral School **Sociétés, politique, santé publique**
University Department **Bordeaux Population Health Research Center (BPH)**

Thesis defended by **Jean Noël Nikiema**
Defended on **10th October, 2019**
In order to become Doctor from Université de Bordeaux

Academic Field **Public Health**
Speciality **Health and informatics**

Thesis Title

Integrating heterogeneous biomedical knowledge through a model based on support ontologies

Thesis supervised by	Fleur Mougin	Supervisor
	Vianney Jouhet	Co-Supervisor

Committee members

Referees	Anita Burgun	Professor at Paris Descartes University France	
	Stefan Schulz	Professor at Medical University Graz Austria	
Examiners	Olivier Bodenreider	Senior Researcher at NLM/NIH USA	
	Geneviève Chene	Professor at Bordeaux University France	Committee President
Supervisors	Fleur Mougin	associate-professor, HDR at Bordeaux University France	
	Vianney Jouhet	PH at CHU of Bordeaux	

Mots clés : intégration sémantique, terminologies biomédicales, ontologies de support

Keywords: semantic integration, biomedical terminology, support ontology

Cette thèse a été préparée au

Bordeaux Population Health
Research Center (BPH)
INSERM U1219, ERIAS
Université de Bordeaux , Case 11
146 rue Léo Saignat
33076 Bordeaux cedex
France
☎ (33) (0)5 57 57 13 93
📠 (33) (0)5 56 24 00 81
Site `https://www.bordeaux-population-health.center/`

A mes fils NIKIEMA Kiswendsida Jean Dyllan et Eliakim Joseph Rayimwende et leur mère Yacine R. OUEDRAOGO. Dyllan, Eliakim vous êtes ma vie, ma source d'inspiration. C'est pour vous que je me bats, c'est grâce à vous que jamais je ne baisserai les bras. Yacine, tu es mon amour et tu le seras à jamais. Puissions-nous réaliser de grands projets ensemble pour le bonheur de tous

A mon père, NIKIEMA Jean-Christophe et ma mère KAFANDO Noélie. Vous avez toujours été un exemple. Votre vision de la vie, votre rigueur et votre affection ont fait de moi l'homme que je suis. Je ne serai jamais assez reconnaissant à votre égard. Mon respect et mon amour pour vous resteront à tout jamais.

A mes frères et soeurs, Rosine, Jacqueline, Christelle, Antoine et Richard

A mes neveux et nièces Marc-Aurèle, Ariel et Camelia

Le diplôme vaut ce que vaut son titulaire, il ne constitue qu'une borne indicatrice amovible, plantée sur la poudre et interminable route du savoir, et qui se déplace vers plus ou moins l'infini selon que l'on continue à s'instruire ou que l'on s'endort sur l'acquis.

Nazi Boni

Productions scientifiques

Articles de revue

NIKIEMA Jean Noël, JOUHET Vianney, MOUGIN Fleur. "Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts". Journal of Biomedical Informatics, 2017, vol. 74, p. 46-58.

NIKIEMA Jean Noël, MOUGIN Fleur, JOUHET Vianney. "Finding the appropriate transversal relations between knowledge resources to semantically enrich their integration process". *Under revision*.

NIKIEMA Jean Noël, GRIFFIER Romain, JOUHET Vianney, MOUGIN Fleur. "Alignment of an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®)". *In preparation*.

Articles de conférences internationales avec publication des actes

NIKIEMA Jean Noël, BODENREIDER Olivier. "Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability", Proceedings of the 10th International Conference on Biomedical Ontology (ICBO), 2019.

NIKIEMA Jean Noël, MOUGIN Fleur, JOUHET Vianney. "Utilisation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie", Proceedings of the 6èmes Journées Francophones sur les Ontologies (JFO), 2016

Articles de conférences nationales

NIKIEMA Jean Noël, JOUHET Vianney, MOUGIN Fleur. “Processus d’intégration de ressources termino-ontologiques en santé”, Proceedings of the IA&Santé workshop at 30èmes journées francophones d’Ingénierie des Connaissances (IC), 2019.

NIKIEMA Jean Noël, MOUGIN Fleur, JOUHET Vianney. “Processus de prétraitement des libellés d’une terminologie d’interface”. Proceedings of the 4ème Symposium sur l’Ingénierie de l’Information Médicale, 95-103, 2017.

NIKIEMA Jean Noël, JOUHET Vianney, MOUGIN Fleur. “Evaluation de la SNOMED CT comme support à l’alignement de terminologies diagnostiques en cancérologie”, Proceedings of the IA&Santé workshop at 27èmes journées francophones d’Ingénierie des Connaissances (IC), 2016.

Intégration de connaissances biomédicales hétérogènes grâce à un modèle basé sur les ontologies de support

Résumé

Dans le domaine de la santé, il existe un nombre très important de sources de connaissances, qui vont de simples terminologies, classifications et vocabulaires contrôlés à des représentations très formelles, que sont les ontologies. Cette hétérogénéité des sources de connaissances pose le problème de l'utilisation secondaire des données, et en particulier de l'exploitation de données hétérogènes dans le cadre de la médecine personnalisée ou translationnelle. En effet, les données à utiliser peuvent être codées par des sources de connaissances décrivant la même notion clinique de manière différente ou décrivant des notions distinctes mais complémentaires. Pour répondre au besoin d'utilisation conjointe des sources de connaissances encodant les données de santé, nous avons étudié trois processus permettant de répondre aux conflits sémantiques (difficultés résultant de leur mise en relation) : (1) l'alignement qui consiste à créer des relations de mappings (équivalence et/ou subsumption) entre les entités des sources de connaissances, (2) l'intégration qui consiste à créer des mappings et à organiser les autres entités dans une même structure commune cohérente et, enfin, (3) l'enrichissement sémantique de l'intégration qui consiste à créer des mappings grâce à des relations transversales en plus de celles d'équivalence et de subsumption. Dans un premier travail, nous avons aligné la terminologie d'interface du laboratoire d'analyses du CHU de Bordeaux à la LOINC. Deux étapes principales ont été mises en place : (i) le prétraitement des libellés de la terminologie locale qui comportaient des troncatures et des abréviations, ce qui a permis de réduire les risques de survenue de conflits de nomenclature, (ii) le filtrage basé sur la structure de la LOINC afin de résoudre les différents conflits de confusion. Deuxièmement, nous avons intégré RxNorm à la sous-partie de la SNOMED CT décrivant les connaissances sur les médicaments afin d'alimenter la SNOMED CT avec les entités de RxNorm. Ainsi, les médicaments dans RxNorm ont été décrits en OWL grâce à leurs éléments définitionnels (substance, unité de mesure, dose, etc.). Nous avons ensuite fusionné cette représentation de RxNorm à la structure de la SNOMED CT, résultant en une nouvelle source de connaissances. Nous avons ensuite comparé les équivalences inférées (entre les entités de RxNorm et celles de la SNOMED CT) grâce à cette nouvelle structure avec les équivalences préétablies de manière morphosyntaxique par RxNorm. Notre méthode a résolu des conflits de nomenclature mais était confrontée à certains conflits de confusion et d'échelle permettant ainsi de mettre en évidence des éléments d'amélioration dans RxNorm et la SNOMED CT. Finalement, nous avons réalisé une intégration sémantiquement enrichie de la CIM10 et de la CIMO3 en utilisant la SNOMED CT comme support. La CIM10 décrivant des diagnostics et la CIMO3 décrivant cette notion suivant deux axes différents (celui des lésions histologiques et celui des localisations anatomiques), nous avons utilisé la structure de la SNOMED CT pour retrouver des relations transversales entre les concepts de la CIM10 et de la CIMO3 (résolution de conflits ouverts). Au cours du processus, la structure de la SNOMED CT a également été utilisée pour supprimer les mappings erronés (conflits de nomenclature et de confusion) et désambiguïser les cas de mappings multiples (conflits d'échelle).

Mots clés : intégration sémantique, terminologies biomédicales, ontologies de support

Bordeaux Population Health – Research Center (BPH)
INSERM U1219, ERIAS – Université de Bordeaux , Case 11 – 146 rue Léo Saignat – 33076 Bordeaux cedex – France

Integrating heterogeneous biomedical knowledge through a model based on support ontologies

Abstract

In the biomedical domain, there are almost as many knowledge resources in health as there are application fields. These knowledge resources, described according to different representation models and for different contexts of use, raise the problem of complexity of their interoperability, especially for actual public health problematics such as personalized medicine, translational medicine and the secondary use of medical data. Indeed, these knowledge resources may represent the same notion in different ways or represent different but complementary notions. For being able to use knowledge resources jointly, we studied three processes, which can overcome semantic conflicts (difficulties encountered when relating distinct knowledge resources): the alignment, the integration and the semantic enrichment of the integration. The alignment consists in creating a set of equivalence or subsumption mappings between entities from knowledge resources. The integration aims not only to find mappings but also to organize all knowledge resource entities into a unique and coherent structure. Finally, the semantic enrichment of integration consists in finding all the required mapping relations between entities of distinct knowledge resources (equivalence, subsumption, transversal and, failing that, disjunction relations). In this frame, we firstly realized the alignment of laboratory tests terminologies: LOINC and the local terminology of Bordeaux hospital. We pre-processed the noisy labels of the local terminology to reduce the risk of naming conflicts. Then, we suppressed erroneous mappings (confounding conflicts) using the structure of LOINC. Secondly, we integrated RxNorm to SNOMED CT. We constructed formal definitions for each entity in RxNorm by using their definitional features (active ingredient, strength, dose form, etc.) according to the design patterns proposed by SNOMED CT. We then integrated the constructed definitions into SNOMED CT. The obtained structured was classified and the inferred equivalences between RxNorm and SNOMED CT were compared to morphosyntactic mappings. Our process resolved some cases of naming conflicts but was confronted to confounding or scaling conflicts highlighting the needs of improvement in RxNorm and SNOMED CT. Finally, we performed a semantically enriched integration of ICD-10 and ICD-O3 using SNOMED CT as support. As ICD-10 describes diagnoses and ICD-O3 describes this notion according to two different axes (i.e., histological lesions and anatomical structures), we used the SNOMED CT structure to identify transversal relations between their entities (resolution of open conflicts). During the process, the structure of the SNOMED CT was also used to suppress erroneous mappings (naming and confusion conflicts) and disambiguate multiple mappings (scale conflicts).

Keywords: semantic integration, biomedical terminology, support ontology

Remerciements

Ce travail est l'aboutissement d'un long cheminement au cours duquel j'ai bénéficié de l'encadrement, des encouragements et du soutien de plusieurs personnes, à qui je tiens à dire profondément et sincèrement merci.

A mes directeurs de thèse Fleur Mougin et Vianney Jouhet, merci d'avoir inspiré ce sujet. Vous m'avez initié et patiemment accompagné dans mes premiers pas dans la recherche en informatique de santé. Vous avez nourri et soutenu en moi le sentiment du travail bien fait. Merci pour vos disponibilités, vos précieux conseils m'ont servi de repère et de guide. Vous avez cru en moi, j'espère avoir été digne de la confiance portée en ma personne. Soyez-en remerciés. Que la grâce surabonde dans votre vie.

A tous les membres de l'équipe ERIAS. Merci à Frantz Thiessard, Gayo Diallo, Valérie Kiewsky, Bruno Thiao-Layel, Aaron Ayllon Benitez, Georgeta Bordea et Marie-Odile Coste vous êtes une équipe formidable qui réalise des choses incroyables dans la bonne humeur et l'entente cordiale. Merci pour ces discussions, ces interrogations passées, ces outils et ce savoir partagés.

Je voudrais dire un merci particulier à Olivier Bodenreider, qui m'a accueilli et formé pendant mon séjour au US NIH. Merci pour sa rigueur, sa soif de la précision, du juste et du beau. Merci de m'avoir fait confiance en m'intégrant dans un de vos projets de recherche. Vous m'avez permis d'obtenir une grande expérience.

Je souhaite remercier toute l'équipe enseignante de l'ISPED, auprès de laquelle j'ai pu effectuer un monitorat pendant deux années.

Merci aux membres du bureau John Snow. Hadrien, Solenne, Irène, Marie, Laura merci pour ces déjeuners passés ensemble, merci pour les débats qui passaient de la “pertinence" du darwinisme à la “magie" des trous noirs, de l'éthique dans la science et la médecine à la géopolitique mondiale. Merci à Quentin, Chloé, Loïc, Henry, Melany, Fabien, Van Hung, Alexandra et Perrine.

Merci à tous les doctorants de l'ISPED, pour ces rencontres du mercredi, ces mots de réconfort et de soutien partagés lors des furtives rencontres dans le couloir.

Je me sens chanceux d'avoir rencontré tant de gens heureux. Je suis privilégié d'avoir pu grandir grâce à vous. Vous avez participé à ma construction. Vous avez une place particulière dans mon cœur tant ces années ont été heureuses grâce à vous.

Merci à Rachid et sa compagne Jessica, Ibrahim et sa compagne Johanna, Thierry et son épouse Amina, Rémi, Sylvie Maurice, Fabrice, Jonathan, Désiré, Jean-Baptiste, Léonse merci pour ces encouragements, ces conseils et tout ce soutien. Vous n'avez jamais été loin quand tout allait bien, vous avez toujours été là quand il le fallait.

Un merci particulier à mes anciens encadrants qui ont toujours su me porter : Pr Macaire Ouedraogo, Pr Drabo Maxime, Pr Robert T. Guiguemde, Pr Abdoul-Salam Ouedraogo et le Pr Nicolas Meda.

Table of contents

> Un concept est une invention à laquelle rien ne correspond exactement, mais à laquelle nombre de choses ressemblent
>
> Friedrich Nietzsche

Substantial summary

Introduction

Dans le domaine de la santé, il existe un nombre très important de sources de connaissances (SCs) [1], qui vont de simples terminologies, classifications et vocabulaires contrôlés à des représentations très formelles, que sont les ontologies [2]. Nous utilisons par la suite le terme de sources de connaissances pour désigner ces différents types de représentation [3].

Les SCs biomédicales constituent un groupe hétérogène puisqu'elles ont été créées avec des niveaux de complexité différents. L'hétérogénéité de ce groupe rend leur interopérabilité complexe [4]. En effet, l'utilisation secondaire des données de santé [5] pour la recherche, la définition de politiques de santé et la médecine personnalisée [6, 7] sont autant de champs nécessitant l'intégration de données de natures diverses, provenant de différents systèmes d'information et codées suivant des SCs différentes. Il est ainsi nécessaire de pouvoir utiliser ces SCs de manière conjointe en ayant une vue complète et cohérente.

Dans la littérature, trois grands types d'hétérogénéité entre SCs ont été recensés [8, 9]. On distingue ainsi :

— l'hétérogénéité **syntaxique** : elle correspond aux différences dues au langage utilisé pour décrire les SCs. Il s'agit de différences dans les formats d'écriture des SCs (Resource Description Framework (RDF [1]), Simple Knowledge Organization System (SKOS [2]), Web Ontology Language (OWL [3]), etc.),

— l'hétérogénéité **structurelle** : elle correspond aux différentes manières de représenter des données dans un même format (modèle d'écriture des termes, type d'organisation hiérarchique des notions, etc.),

— l'hétérogénéité **sémantique** : elle correspond aux différences dans les notions représentées (maladies, processus biologiques, actes médicaux, actes

[1] https ://www.w3.org/RDF/

[2] https ://www.w3.org/TR/skos-reference/

[3] https ://www.w3.org/TR/owl-features/

infirmiers, etc.).

Ce travail de thèse décrit l'intérêt d'utiliser une SC de support pour garantir l'interopérabilité sémantique entre SCs. La première section présente les différents processus applicables pour surmonter les hétérogénéités entre SCs. Les trois sections suivantes introduisent les techniques que nous avons implémentées pour mettre en œuvre les processus précédemment identifiés. Nous discutons finalement l'intérêt d'utiliser une SC de support tel que cela a été mis en évidence dans chaque processus.

Cadre d'étude

Les correspondances entre entités de deux SCs peuvent être retrouvées manuellement par des experts du domaine. Cependant, les SCs pouvant contenir une grande quantité d'entités, il s'agit d'un processus qui peut être long et fastidieux. L'alternative est d'établir les correspondances de manière automatique par la création de **mappings**, qui consiste à déterminer une expression formelle de la relation sémantique entre deux entités. Un mapping est souvent représenté par un quintuplet <id, e_1, e_2, r, n>, où e_1 et e_2 sont les deux entités à lier, *id* est l'identifiant de la correspondance, *r* la relation sémantique entre les deux entités et *n* la mesure de confiance associée [10]. Deux processus existent pour établir des correspondances entre des SCs :

— l'**alignement** qui vise à créer des mappings entre les entités des différentes SCs [11],

— l'**intégration** qui consiste à créer une nouvelle SC en utilisant des SCs préexistantes [12]. En pratique, cela nécessite d'établir des mappings entre les entités des SCs à intégrer, puis à réorganiser les SCs dans une structure conjointe et unique.

Dans la littérature, ce sont essentiellement les relations d'**équivalence** et de **subsomption** qui sont trouvées lors de l'identification des mappings. Les entités représentant des notions différentes sont au mieux associées via des relations de **disjonction**. Ainsi, quand les notions entre les SCs à relier sont distinctes mais complémentaires, les solutions proposées sont insuffisantes. Les travaux existants se contentent d'organiser de manière cohérente les entités représentant des notions différentes mais ne créent pas de liens directs entre les entités en cas de complémentarité [13]. Pour répondre à cette problématique, nous introduisons le besoin d'enrichir sémantiquement le processus d'intégration. Ce processus vise à créer des mappings (relations d'équivalence ou de subsomption), à organiser de manière cohérente les entités, puis à identifier des relations transversales entre les entités qui sont différentes mais complémentaires (*e.g.*, mappings entre gènes et médicaments, entre maladies et données géographiques). Il repose sur deux étapes : (1) l’ancrage à une SC de support, et (2) la dérivation suivant la SC de support.

Processus d’alignement : cas des analyses biologiques

La LOINC® (Logical Observation Identifiers Names and Codes [14, 15]) est une SC de support dans le domaine des analyses biologiques. Procéder à l’alignement de SCs locales avec la LOINC permet d’assurer l’utilisation conjointe des données de biologie provenant de plusieurs structures de soins. Ainsi, en alignant les entités des SCs locales avec celles de la LOINC, les entités LOINC peuvent être utilisées pour obtenir des données comparables à travers différents systèmes d’information.

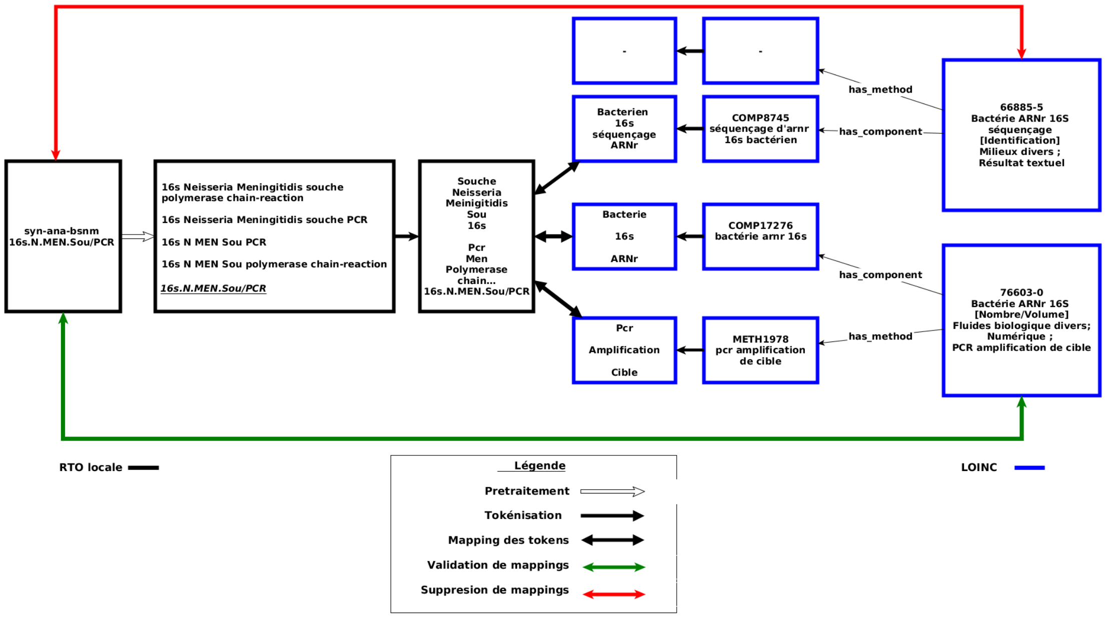

syn-ana-bsnm
16s.N.MEN.Sou/PCR
16s Neisseria Meningitidis souche polymerase chain-reaction
16s Neisseria Meningitidis souche PCR
16s N MEN Sou PCR
16s N MEN Sou polymerase chain-reaction
16s.N.MEN.Sou/PCR
Souche Neisseria Meinigitidis Sou 16s
Pcr Men Polymerase chain... 16s.N.MEN.Sou/PCR
-
-
Bacterien 16s séquençage ARNr
COMP8745 séquençage d'arnr 16s bactérien
Bacterie 16s ARNr
COMP17276 bactérie arnr 16s
Pcr Amplification Cible
METH1978 pcr amplification de cible
has_method
has_component
has_component
has_method
66885-5 Bactérie ARNr 16S séquençage [Identification] Milieux divers ; Résultat textuel
76603-0 Bactérie ARNr 16S [Nombre/Volume] Fluides biologique divers; Numérique ; PCR amplification de cible
RTO locale
Légende
Pretraitement
Tokénisation
Mapping des tokens
Validation de mappings
Suppresion de mappings
LOINC

Figure 1 – Exemple d’alignement d’un concept de la SC locale du CHU de Bordeaux à un concept LOINC.

Nous avons aligné la SC locale du CHU de Bordeaux à la LOINC. La méthodologie reposait sur trois étapes. La première étape a consisté au pré-traitement des libellés de la SC locale [16]. La deuxième étape visait à calculer la similarité morphosyntaxique entre les tokens constitutifs des libellés de la SC locale et de la LOINC. Dans la troisième étape, nous avons utilisé la structure de la LOINC

pour procéder au filtrage des mappings obtenus (Figure 1). En effet, les éléments définitionnels des concepts de la LOINC sont délimités dans chacun des libellés grâce à une ponctuation précise. Le caractère ":" sépare les concepts LOINC en leurs éléments principaux, comme suit :
<composant/analyte>:<propriété>:<temps>:<milieu_biologique>:
<échelle>:<méthode>.
Ainsi, comme déjà implémenté dans [17], nous avons créé des relations entre le concept LOINC et chacun de ses éléments définitionnels. Chaque relation a été nommée en combinant le préfixe *has_* et le type d'élément définitionnel (*has_component*, *has_property*, etc.).

Nous avons utilisé ServoMap [18] pour créer des mappings entre les concepts de la SC locale et les éléments définitionnels des concepts LOINC. Ensuite, des relations de mapping ont été créées entre les concepts de la LOINC et ceux de la SC locale qui partageaient le même analyte. Ces mappings ont ensuite été filtrés grâce à la structure de la LOINC par la suppression de mappings erronés, à savoir des mappings entre concepts n'appartenant pas au même chapitre et ne décrivant pas le même milieu biologique ou la même méthode. Dans ce processus d'alignement, la structure de la LOINC a donc permis de procéder à la correction de mappings, palliant ainsi les limites de la SC locale.

Processus d'intégration : cas de la représentation du médicament

Dans un deuxième travail, nous avons intégré RxNorm [15] à la sous-partie de la SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms) [15, 19] décrivant les connaissances sur les médicaments afin d'alimenter la SNOMED CT avec les entités de RxNorm. La SNOMED CT étant une référence au niveau international et RxNorm étant utilisée aux États Unis, procéder à l'intégration de ces deux SC vise à rendre interopérables les données sur le médicament présentes dans les systèmes d'information sanitaire (SIS) aux États Unis à n'importe quel SIS dans le monde qui utilise la SNOMED CT. De plus, un nouveau modèle de représentation du médicament a été décrit au sein de la SNOMED CT [19]. Ce modèle étant basé sur les recommandations internationales regroupées au sein de l'IDMP (Identification of Medicinal Products) [20], l'intégration permettra d'évaluer la conformité de RxNorm aux règles internationales de description du médicament.

Nous avons représenté les concepts RxNorm selon le modèle de description du médicament utilisé par la SNOMED CT [19]. Ainsi, les médicaments dans RxNorm ont été décrits en OWL grâce à leurs éléments définitionnels (substance, unité de mesure, dose, etc.). Nous avons ensuite fusionné cette représentation de RxNorm à la structure de la SNOMED CT, résultant en une nouvelle SC composée de RxNorm et de la SNOMED CT. Au sein de cette nouvelle SC, nous avons créé des mappings d'équivalence entre les éléments définitionnels des entités de RxNorm et celles de la SNOMED CT. Enfin, nous avons généré la structure inférée de cette nouvelle SC en utilisant le raisonneur ELK [21, 22]. Ce choix est motivé par le fait que ce raisonneur a été décrit comme étant le plus adapté pour classer la SNOMED CT [23].

Nous avons comparé les “mappings déclarés” (*i.e.*, les mappings créés de manière morphosyntaxique par les concepteurs de RxNorm entre les médicaments dans RxNorm et ceux dans la SNOMED CT) aux “mappings inférés” (mappings obtenus par la classification de la structure de la nouvelle SC réalisée par le raisonneur).

Le tableau 1 décrit la distribution des concepts SNOMED CT en fonction de leur mappings aux concepts de RxNorm.

Tableau 1 – Distribution des concepts SNOMED CT décrivant le médicament, en fonction de leur mapping aux concepts RxNorm: comparaison des mappings inférés aux mappings définis par RxNorm (mappings déclarés)

		Mappings déclarés		**Total**
		Présents	Absents	
Mappings inférés	Présents	1 892	110	2 002
	Absents	939	263	1 202
Total		2 831	398	3 204

L'intégration de RxNorm et de la SNOMED CT a permis de mettre en évidence l'intérêt d'exploiter les éléments définitionnels des concepts de chaque SC. Certains mappings ont pu être retrouvés uniquement de manière inférée (110 mappings), témoignant de limites potentielles de la méthode morphosyntaxique utilisée par les concepteurs de RxNorm. En revanche, 939 mappings déclarés n'ont pas été retrouvés par le processus que nous avons mis en œuvre. Si certaines différences étaient consécutives aux limites de notre processus d'intégration (telles que l'absence de conversion de certaines unités de mesure. Pour ce cas spécifique, une solution pourra être implémentée en utilisant l'UCUM[4] (Unified Code for Units of Measure)). Nous avons par ailleurs identifié des incohérences dans les éléments définitionnels des concepts. En effet, il existe des différences de précision dans la définition de certains concepts. Par exemple, RxNorm n'utilise pas d'“unité de présentation” (unité comptable dans laquelle les médicaments sont présentés) pour décrire ses entités, contrairement à la

[4] https://unitsofmeasure.org/trac

SNOMED CT. Inversement, RxNorm utilise des “Qualitative Distinction” (*i.e.*, des étiquettes qui sont cliniquement pertinentes telles que ”sans sucre”) pour la description des médicaments, ce que ne fait pas la SNOMED CT. Des différences dans des éléments majeurs tels que le “Basis of strength” (substance de référence de la dose du médicament) ont également été trouvées, ce qui change fondamentalement la définition. La mise en évidence de ces différences a permis de proposer des pistes d’amélioration de la description du médicament dans les deux SCs [24].

Enrichissement sémantique d’un processus d’intégration: cas de la cancérologie

Dans le domaine de la cancérologie, la réutilisation des données est confrontée à l’hétérogénéité des SCs utilisées pour le codage des diagnostics. Afin de pallier cette difficulté, il est nécessaire de mettre en correspondance ces différentes SCs et, en particulier, la CIM-10 (dixième révision de la Classification statistique Internationale des Maladies et des problèmes de santé connexes [25]) et la CIM-O3 (la troisième révision de la Classification Internationale des Maladies pour l’Oncologie [26]). Ces deux SCs sont utilisées de manière différente pour coder des diagnostics : la CIM-10 les décrit en tant que tels tandis que la CIM-O3 décrit des lésions histologiques et des localisations anatomiques, qui sont représentées suivant deux axes distincts et peuvent être combinées. Les notions représentées par ces SCs étant distinctes, nous avons réalisé un processus d’intégration semantiquement enrichi entre elles en utilisant la SNOMED CT comme support.

Pour cela, deux étapes ont été mises en œuvre. Premièrement, la CIM-10 et la CIM-O3 ont été alignées à la SNOMED CT. Cette étape d’alignement, qualifiée d’**ancrage**, vise à rechercher des mappings d’équivalence entre les concepts de la CIM-10 et de la CIM-O3 et ceux de la SNOMED CT. La structure de la SNOMED CT a servi à : (1) filtrer les mappings incorrects (notamment les mappings entre des concepts de maladie et d’anatomie), et (2) désambiguïser les mappings multiples. La deuxième étape, dite de **dérivation**, a consisté à établir des mappings complexes entre un concept CIM-10 et une paire de concepts CIM-O3. Tout d’abord, nous avons cherché dans la SNOMED CT les relations transversales pouvant lier les concepts de la CIM-10 et ceux de la CIM-O3. Nous avons ainsi identifié *finding_site* pour associer les diagnostics et les localisations anatomiques et *associated_morphology* entre les diagnostics et les lésions histologiques. Nous avons ensuite construit une structure inférée de la SNOMED CT grâce au raisonneur ELK [22, 23]. Sur la base des relations transversales iden-

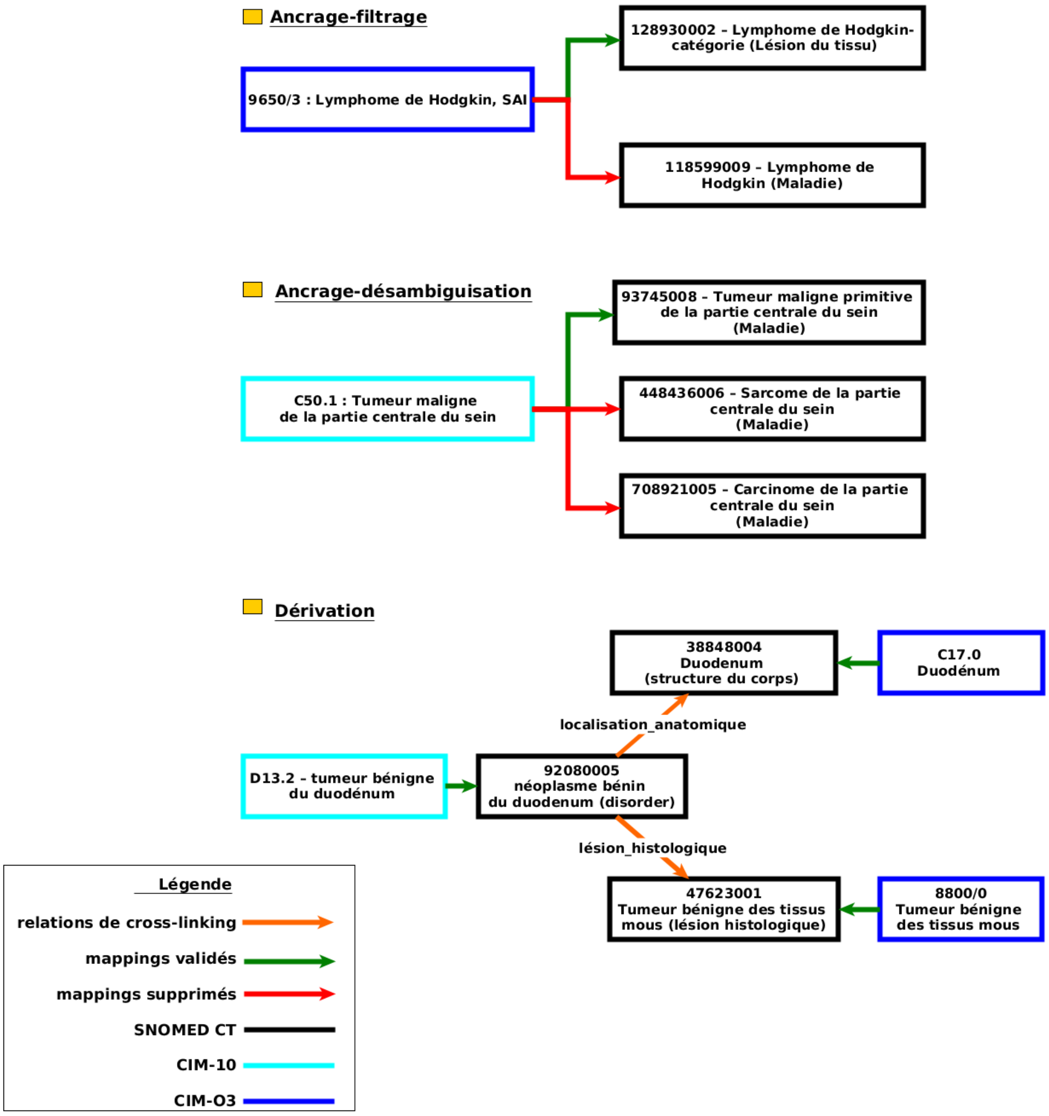

Ancrage-filtrage
9650/3 : Lymphome de Hodgkin, SAI
128930002 - Lymphome de Hodgkin-catégorie (Lésion du tissu)
118599009 - Lymphome de Hodgkin (Maladie)
Ancrage-désambiguisation
C50.1 : Tumeur maligne de la partie centrale du sein
93745008 - Tumeur maligne primitive de la partie centrale du sein (Maladie)
448436006 - Sarcome de la partie centrale du sein (Maladie)
708921005 - Carcinome de la partie centrale du sein (Maladie)
Dérivation
38848004 Duodenum (structure du corps)
C17.0 Duodénum
localisation_anatomique
D13.2 - tumeur bénigne du duodénum
92080005 néoplasme bénin du duodenum (disorder)
lésion_histologique
47623001 Tumeur bénigne des tissus mous (lésion histologique)
8800/0 Tumeur bénigne des tissus mous
Légende
relations de cross-linking
mappings validés
mappings supprimés
SNOMED CT
CIM-10
CIM-O3

Figure 2 – Exemple d'enrichissement sémantique de l'intégration basé sur les relations transversales décrites dans la SNOMED CT

tifiées, nous avons repéré les concepts CIM-10 équivalents à une combinaison de concepts CIM-O3 de morphologie et de topographie. Notons que l'identification d'inférences erronées dans l'étape de dérivation a permis de détecter des inconsistances dans la SNOMED CT.

L'enrichissement sémantique du processus d'intégration résulte en ce qui a été appelé dans la littérature des mappings complexes [27]. On parle de mapping complexe quand la relation de mapping est établie entre deux éléments dont au moins un des éléments n'est pas une simple entité. Ainsi, pour aller plus loin que l'établissement de mappings simples entre un code CIM-10 lié à un code CIM-O3 de topographie par une relation transversale, nous avons identifié des relations d'équivalence (lorsque le concept SNOMED CT était défini) ou de subsomption (lorsque le concept SNOMED CT était primitif) entre un concept CIM-10 et un couple de code CIM-O3. Nous avons automatiquement dérivé 86% (892/1032) des concepts CIM-O3 morphologiques avec 38% (127/330) de concepts CIM-O3 topographiques et 24% (203/852) des concepts CIM-10. La dérivation, analysée manuellement, a permis d'identifier des erreurs dans la hiérarchie de la SNOMED CT. Par exemple, elle a mis en évidence une relation de subsomption erronée entre les concepts 20955008-*insulinome malin* et 3898006-*néoplasme bénin* (version de janvier 2017). Cette erreur a depuis été corrigée lors de la mise à jour de la SNOMED CT.

En conclusion, l'enrichissement sémantique du processus d'intégration a permis de mettre en évidence l'intérêt d'utiliser une SC de support pour les tâches suivantes : (i) la correction de mappings erronés lors de la phase d'ancrage, (ii) la découverte de mappings impliquant des relations transversales, et (iii) l'audit indirect de la SC de support lorsque des inférences erronées ont été identifiées.

Intérêts de l'utilisation d'une source de connaissances de support

Pour l'alignement et l'intégration, les stratégies appliquées dans notre travail et dans la littérature se basent essentiellement sur le calcul de mesures de similarité entre entités provenant de différentes SCs. Ces mesures de similarité sont généralement calculées d'après des techniques morphosyntaxiques, structurelles et sémantiques. Il est important de souligner que les similarités obtenues peuvent donner lieu à des interprétations erronées : ce sont les **conflits sémantiques** [28, 29]. Or, les conflits sémantiques ne sont pas tous résolus par les processus d'alignement et d'intégration.

Les méthodes morphosyntaxiques, consistant à retrouver des similarités entre les libellés des entités, sont souvent utilisées en premier lors de la création automatique de mappings [30, 31]. Ces méthodes sont confrontées au risque de survenue de **conflits de nomenclature**, qui sont consécutifs aux similarités ou dissimilarités incorrectes entre des termes utilisés pour désigner les entités des SCs. Ainsi :

— dans les cas d'homonymie, des mappings sont établis de manière erronée entre des concepts différents. Cette situation a été illustrée par les mappings qu'il était nécessaire de filtrer entre les concepts de la SC locale et ceux de la LOINC.

— dans les cas de synonymie, certains concepts pourtant équivalents ne sont pas mappés.

Les méthodes structurelles sont habituellement utilisées après les méthodes morphosyntaxiques. Elles consistent à calculer le niveau de chevauchement des instances ou la proximité taxonomique des concepts présents dans les SCs. Ces stratégies peuvent résoudre des cas de synonymie [30]. Ainsi, à partir de la structure de la SNOMED CT et de RxNorm, des mappings ont pu être établis entre des concepts qui n'avaient pas été mappés par des méthodes morphosyntaxiques (Tableau 1). Cependant, les méthodes structurelles étant tributaires de la qualité de la structure des SCs, elles sont sujettes aux conflits d'échelle et de confusion. Les **conflits d'échelle** apparaissent lorsqu'il y a une différence de granularité dans les définitions des notions représentées (*e.g.*, absence d'"unités de présentation" dans la définition des concepts de RxNorm). Les **conflits de confusion** sont dus à des définitions contradictoires (*e.g.*, les différences de "Basis of strength" entre les concepts RxNorm et ceux de la SNOMED CT).

Les méthodes sémantiques décrites dans la littérature consistent à utiliser un support de connaissances. A partir des mappings avec une SC de support, les entités de celles-ci servent à établir des ponts entre les SCs à mettre en correspondance. Par exemple, à partir d'un concept de l'UMLS (Unified Medical Language System), il est possible de retrouver tous les concepts des SCs intégrées dans l'UMLS qui sont censées décrire la même notion [32–34]. Ainsi, les stratégies proposées dans la littérature permettent de réaliser l'alignement ou l'intégration de SCs. Néanmoins, parce que ces stratégies se limitent à la recherche d'entités équivalentes ou reliées hiérarchiquement entre différentes SCs, il n'est pas possible de les utiliser pour relier des entités décrivant des notions différentes mais complémentaires. Le processus permettant de prendre en compte ces limites est ce que nous avons qualifié l'enrichissement sémantique du processus d'intégration. Celui-ci est basé sur une méthode combinant les méthodes de calcul de similarité pour répondre aux problématiques de conflits de confusion et d'échelle, tout en établissant des correspondances entre des entités différentes via des relations transversales [35, 36]. Cette méthode repose sur l'utilisation d'une SC de support et consiste en deux étapes : l'ancrage et la dérivation.

Comme les processus d'alignement et d'intégration que nous avons mis en place, l'ancrage à une SC formelle permet la mise en place de procédures de validation des mappings basée sur la structure de ce support. La SC de support doit disposer d'une structure formelle et être, au mieux, une ontologie pour une stratégie de mise en correspondance optimale. Les SCs de support apportent des éléments définitionnels aux entités participant aux mappings, ce qui permet de s'affranchir de la qualité des structures des SCs à relier. En effet, l'ancrage apporte des synonymes [30] et supprime des mappings erronés [37] (comme illustré dans la figure 2 avec le filtrage en cas de conflit de confusion et la désambiguisation en cas de conflit d'échelle). Dans l'enrichissement des processus d'integration, la dérivation est l'étape essentielle qui permet d'améliorer l'organisation entre SCs en reliant les entités différentes par des relations transversales grâce à la structure de la SC de support.

Conclusion

Notre étude présente trois processus permettant d'utiliser conjointement des SCs biomédicales hétérogènes. Deux aspects résument l'intérêt d'exploiter une SC de support dans ce cadre : (1) la résolution des différents conflits sémantiques en procédant au filtrage de mappings erronés et à la désambiguïsation de mappings multiples, et (2) la possibilité d'établir automatiquement des mappings complexes entre des entités différentes mais décrivant des notions complémentaires.

...language... reflects our obsessions and ways of conceptualizing the world.

Steven Pinker

Chapter 1

Integrating knowledge resources: challenges and future trends

Summary: In this chapter, we describe the necessity to semantically enrich the integration process of knowledge resources. More specifically, this chapter motivates the need to reveal all possible links between knowledge resources, even if they are different in their structure.

Thus, we described how mappings can be created on the basis of different types of relation: equivalence, subsumption, disjunction and specific transversal relations. By analyzing the literature according to the heterogeneities of knowledge resources, we found that using support ontologies is the appropriate way to find all the appropriate mapping relations between different knowledge resources regardless of their structure.

Indeed, because they exhibit a formal structure, ontologies can be used to create equivalence or hierarchical mappings between knowledge resources even if they are poorly structured. It is most noteworthy that ontologies containing relationships which may associate entities from distinct hierarchies, they are useful to automatically establish mappings based on transversal relations.

Thus, we present different strategies that can be used for integrating knowledge resources and how these strategies differently address the difficulties faced when trying to relate concepts of distinct knowledge resources.

Keywords knowledge resources, semantic integration, support ontologies, semantic conflicts

Valorization This chapter is based on the article entitled “Finding the appropriate transversal relations between knowledge resources to semantically enrich their integration process", which has been submitted for publication in the journal of biomedical semantic (JBS). This article is currently under review.

1.1 Introduction

In the medical domain, knowledge resources are science products that aim at listing the concepts, corresponding to units of knowledge in a given domain [38, 39], and the appropriate terms to refer to them [39–41]. The identification and the naming of domain concepts are tasks performed according to particular **contexts of use** (international reference knowledge resources, local knowledge resources, etc.) and according to predefined **objectives** (clinical use, epidemiological use, bibliographic research etc.). Thus, organizing and naming the inventoried concepts can be done with different levels of complexity.

Whatever the language of description for **naming** these concepts (English, Chinese, French...), terms used to designate them can be written according to a predefined writing convention. This is the case of SNOMED CT® (Systematized Nomenclature of Medicine– Clinical Terms) in which terms have a *"semantic tag" in parenthesis which identifies the hierarchy to which the SNOMED CT-concept belongs* [42], and LOINC® (Logical Observation Identifier Names and Codes) [43] in which punctuation within terms separate them into different sub-parts. Conversely, terms may have acronyms and word truncations that do not meet any basic format. This last characteristic occurs more often in interface terminologies [44]. When **organizing** concepts in a knowledge resource, the existing relationships between these concepts can be represented with different levels of complexity. Thus, knowledge resources can be organized according to hierarchical (like in taxonomies) and/or non-hierarchical (like in classifications) relations. Hence, knowledge resources vary from terminologies, classifications and controlled vocabularies to very formal representations [45]. When knowledge resources are represented in a language that can be operationalized, with entities that are logically defined, the term "ontologies" is used to designate them [2]. Knowledge resources may be built according to different philosophies [46] but they may be used together, precisely in the biomedical domain, regardless of the complexity of their representation. In addition, what is called an "ontology" in the biomedical field is not always conformed to all the characteristics to be named as such [47]. For better readability, throughout the rest of the document, we used the following terms: (i) **knowledge resource** to designate this heterogeneous group, (ii) **ontology** for formal resources, and (iii) **terminology** for non-formal resources.

In this chapter, we firstly described the necessity to overcome the heterogeneity of knowledge resources and relate them. Secondly, we describe the difficulties to relate knowledge resources. These difficulties are described on the basis of the heterogeneity of knowledge resources. Finally, we describe the techniques that can be applied to overcome each difficulty.

1.2 Challenges in finding correspondences between knowledge resources

1.2.1 Needs for relating knowledge resources

Nowadays, the resolution of many health issues deals with the necessity to use jointly data coming from information systems that employ different knowledge resources for data recording.

Indeed, the secondary use of medical data [5], translational medicine [48], personalized medicine [6, 7] and "One health" [49] necessitate to integrate information coming from various systems. These data are often annotated using multiple heterogeneous knowledge resources. Finding semantic correspondences between the entities of distinct knowledge resources that describe these data is thus a requirement.

Secondary use of biomedical data.

Secondary use of biomedical data is a major issue because it supports the improvement of health systems and a better understanding of diseases and treatments. Indeed, it consists in using patient information collected during care delivery for research and billing purposes as well as for certification and accreditation of health facilities, evidence-based medicine and business applications [50].

Secondary use hence opens perspectives for applying data mining approaches to the biomedical domain. These approaches are promising to "*greatly expand the capacity to generate new knowledge*" and "*help translate personalized medicine initiatives into clinical practice by offering the opportunity to use analytical capabilities that can integrate systems biology (e.g., genomics) with electronic health record (EHR) data*" [51]. For example, through the identification of cancer cases by registries, oncology is an area where the secondary use of health data is particularly important [52]. Indeed, the goal of cancer registries is to track all cases of cancer occurring in a defined population. The cancer data are continuously and systematically collected from various healthcare facilities (*e.g.*, hospital, pathology laboratory). The collected data are sociodemographic information about the patients, as well as clinical and histopathological characteristics of the cancer that is being studied. Cancer registries thus allow the follow-up of patients diagnosed with cancer and provide statistical results on the outcomes of the corresponding disease (*e.g.*, mortality, results of therapy) [53]. The collected data are also used for epidemiological research on cancer incidence and determinants, as well as for supporting evidence for health policies on diagnosis, prevention and cancer treatment [54]. Therefore, cancer monitoring requires the concomitant use of data coming from different sources with the difficulty that these data are potentially encoded according to different knowledge resources [55]. Indeed, for encoding data, cancer registries use the third edition of the International Classification of Diseases for Oncology (ICD-O3) [26], while, in France for instance, the medical data in hospitals are encoded (for billing purpose through the PMSI "Programme de Médicalisation du Système d'Information" [56]) according to the tenth revision of the International Statistical Classification of Diseases and related health problems (ICD-10). In addition, morbidity and mortality causes are internationally recorded using ICD-10 [25]. The achievement of the objectives of cancer registries thus requires, among other things, the joint use of ICD-10 and ICD-O3.

One health, translational and personalized medicines.

Secondary use of medical data focuses on reorienting medical data to a purpose other than care. In contrast, translational and personalized medicine as well as "One Health" paradigms are mainly lead by the need to better use all accessible information, even if they have not been created for the purpose of human health care, to make care decisions or disease prevention.

One health mainly consists in integrating information from animal and human medicines [1], while translational medicine consists in using laboratory research information into daily clinical care [2]. Personalized medicine (also designated as genomics medicine or precision medicine) is differently defined in the literature. The US National Institute of Health (NIH) characterized it as "*an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person.*" [3] For Jain *et al.* [57], "*Personalized medicine simply means the prescription of specific therapeutics best suited for an individual based on pharmacogenetic and pharmacogenomic information.*" Thus, personalized medicine requires the integration of genetic, environmental and medicinal product data to improve the treatment of a single person, even if for now, the goals are to deal with sub-groups. As described in [58], knowledge resources can help to prioritize the genes involved in a biological pathway on which the drug has an influence. In this regard, some works

[1] https://www.who.int/features/qa/one-health/en/

[2] https://www.eupati.eu/non-clinical-studies/translational-medicine/

[3] https://ghr.nlm.nih.gov/primer/precisionmedicine/definition

that aim to integrate knowledge resources representing drugs and biological pathways have been realized *in silico* to orient clinical trial approaches [59].

1.2.2 Challenges in overcoming heterogeneities of knowledge resources

The previous section illustrates that many application fields need to use jointly health-related data. Nevertheless, this use is further complicated by the heterogeneity of the knowledge resources used for encoding such data.

Heterogeneity of knowledge resources.

In the literature, the heterogeneity of knowledge resources corresponds to one of the following three types [8, 9, 60–62]:

— **syntactic** heterogeneity: this situation occurs when there is a difference in the writing formats used to describe the knowledge resources (*e.g.*, Resource Description Framework (RDF [4]), Simple Knowledge Organization System (SKOS [5]), Web Ontology Language (OWL [6])),

— **structural** heterogeneity: this second type of heterogeneity appears when there is a difference in the representation of the concepts (*e.g.*, compositional structure of terms, hierarchical and non-hierarchical organization of listed concepts),

— **semantic** heterogeneity: this heterogeneity appears when there is a difference in the knowledge conveyed by the resources (*i.e.*, the represented concepts are different).

Processes for linking knowledge resources.

To overcome the heterogeneity of knowledge resources and to link them, multiple processes exist.

"Translation" [11, 63] (or "morphing" [64]) is commonly used as a first step to link two knowledge resources. This process aims to overcome the syntactic heterogeneity between knowledge resources. It involves either creating a new writing format [65] or using a pre-existing one [11, 66]. Rewriting the format of knowledge resources is essential and has been covered by many publications [64, 67–69]. In the remaining part of our work, we assume that knowledge resources

[4] https://www.w3.org/RDF/

[5] https://www.w3.org/TR/skos-reference/

[6] https://www.w3.org/TR/owl-features/

are written in an appropriate common format and we focus our study on the next steps.

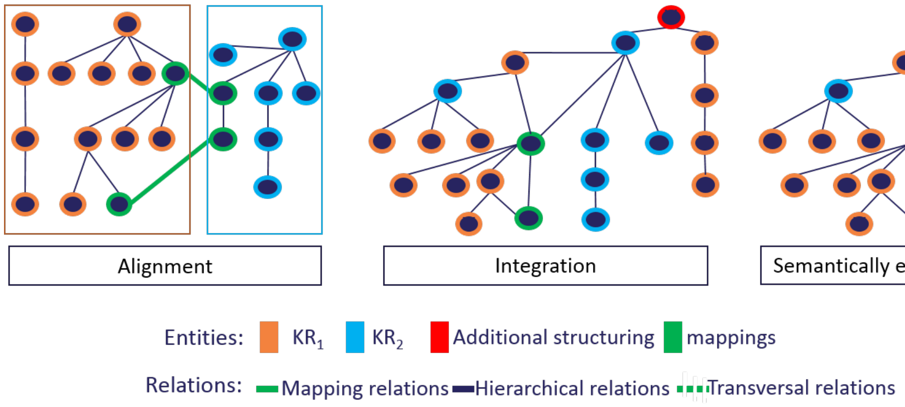


Figure 1.1 – Different processes of semantic integration. The processes are ordered by complexity (from left to right). **Alignment** consists in finding mappings between entities of knowledge resources. **Integration** consists of two steps: finding mappings and organizing all the entities of the resources to be integrated into a unique coherent structure [13]. **Semantically enriched integration** goes beyond the integration ; it consists in additionally identifying the appropriate transversal relations between entities representing different but complementary concepts. KR means "knowledge resource".

Figure 1.1 presents processes that can be implemented to overcome the other types of heterogeneity. For linking knowledge resources, two main strategies may be implemented and combined:

- **Correspondences are established manually by domain experts**. Tools have been developed to support the manual creation of mappings. Examples of manual approaches followed in the biomedical domain include the work of Giannangelo *et al.*, who created mappings between concepts of ICD-10 and SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) [70], and the work of Souvignet *et al.*, who established mappings between concepts and relations of PS-CAST (Patient Safety Categorical Structure), which is an ontology made by the WHO (World Health Organization) and BFO (Basic Formal Ontology), an upper level ontology [71]. However, this task can be tedious and time-consuming;
- **Correspondences are established automatically by creating mappings** [63]. Our work is focused on this task.

A mapping is often presented as a quintuple <id, e_1, e_2, r, n>, where e_1 and e_2 represent the entities to match, *id* the correspondence identifier, *n* the confidence measure and r the semantic relation between e_1 and e_2 [10]. If the semantic relation can be of any type, most of the existing works aim to establish **equivalence** and **subsumption** relations. Such mappings address the issue of the **alignment of knowledge resources** (*e.g.*, in [72–75]), which requires an overlap between concepts represented in the knowledge resources [31].

Mappings are generally established according to **similarity measures** that are calculated between entities [76–79]. When entities cannot be related through equivalence and subsumption relations (*i.e.*, when the mapping algorithm cannot state that the entities are similar), the entities are at most related through a **disjunction** relation. Alignment being defined as a set of correspondences, only the set of mappings corresponding to equivalent or hierarchically related entities are presented.

The step forward, which takes into account the difference of concepts in knowledge resources, is the **integration of knowledge resources**. This process is beyond alignment [76]. There is a misuse of the term "integration" in the literature [12]. Indeed, knowledge resource integration consists in the construction of a new resource by reusing (assembling, specializing and/or adapting) pre-existing ones. These resources may not describe the same domain. In this case, it is possible to identify, in the newly constructed resource, modules that can correspond to the knowledge resources used for integration. Nevertheless, integration is sometimes assimilated to the "merging" process [80] where knowledge have just a different level of granularity. In this merging process, it is difficult to find modules in the new resource that can correspond to the knowledge resources to be merged [12]. In our work, the integration process include the merging process. The resulting knowledge resource is obtained through bridges that can be used to link the entities of the resources to be integrated. These bridges are entities that describe equivalent or hierarchically related concepts through the different resources.

However, the integration process mainly relies on hierarchical relationships available within the resources, thus ignoring the entities that could be related by non-hierarchical or non-equivalence relationships (namely **transversal**) [81–84]. For example, a "breast cancer (disease)" arises in the "breast (organ)". In the mappings, such a disjunction relation must be specified with a more relevant semantic relation, when it is possible. Indeed, linking data that describe different concepts with complementary characteristics can help to better understand diseases and improve treatments at the individual and population levels (*e.g.*, genetic and disease relations, disease and geographic information). Nevertheless, in the literature, when they are related, the entities describing complementary concepts are represented in some simple matrix tables expressing the proximity between them without a relation specifically defined. Although efforts have been made to represent relationships between entities describing complementary concepts, the relation is limited to a general description (*e.g.*, "mapping relationship") [85, 86] or to a created similarity relation [87] that describes a proximity between entities that display different but complementary concepts. Certainly, the use of a general relation between entities may be explained by the fact that these relations are still to be specified in the state-of-the-art. This explanation is not always applicable,though.

As a result of previous assertions, we can conclude that the integration must allow the establishment of mappings that can contain all the relevant types of relation (equivalence, subsumption, transversal and, failing that, disjunction). This statement is the way to induce that a given entity conserves the same meaning across the knowledge resources. Otherwise, we are exposed to the occurrence of **semantic conflicts**.

In the next sections of this chapter, we use the relevant literature as a starting point for describing the semantic conflicts and the features needed to address all of them to semantically enrich the integration process. We firstly propose a general framework to identify which semantic conflicts have already been addressed and which ones need to be explored further.

1.3 Framework to semantically enrich the integration process

1.3.1 The current situation

The semantic conflicts appear when a given information does not have the same meaning across two contexts [88]. Such conflicts lead to difficulties in reconciling the meaning of entities that are described in different knowledge resources [89, 90]. To be resolved, semantic conflicts must be identified and detected [77].

In the literature, based on the interpretation of similarity measures computed between knowledge resources' entities [77–79], the following three types of semantic conflicts [91–93] have been identified:

- **naming conflicts**: irrelevant similarity or dissimilarity in names of entities,
- **scaling conflicts**: different levels of precision in the definitions of entities,
- **confounding conflicts**: contradictory definitions.

Thus, in the literature, the inventoried types of semantic conflicts are named and described on the basis of the interpretations of similarity measures that are computed between entities of knowledge resources [77–79]. Thereby, semantic conflicts are mainly described according to the relevance of the declaration of similarity between two entities [91]. Consequently, these semantic conflicts are not specified by taking into account the entities describing complementary concepts.

The similarity and dissimilarity between entities must be better understood. For this purpose, we firstly need to highlight the characteristics of entities that can be used for computing similarity measures. This idea is to provide a more complete definition of semantic conflicts through the refinement of similarity, but also dissimilarity interpretations to allow a description of semantic relation between entities that represent different but complementary concepts.

1.3.2 Proposed formalism to describe contents of knowledge resources

To formally describe the entities in knowledge resources, we use the semiotic triad of Peirce [94–97]. This triad is composed of:

— the **referent**, corresponding to the object or the knowledge to be represented,

— the **representamen**, corresponding to the sign used to designate the referent,

— the **interpretant**, corresponding to the described meaning of the referent. In [98], it has been defined as *"a meaning of a sign and also another sign explaining the former one"*.

Viewed through the prism of the semiotic triad, each entity in a knowledge resource can be described according to a <R,T,D> triplet as follows:

— R (concept): the referent is the referred concept[7] of an entity and/or a definition in the knowledge resource,

— T (term): the representamen is each element which can be used to designate a concept in a knowledge resource. It may be a label (L) and/or a code or index (I),

— D (description): the interpretant corresponds to the different descriptions of a concept that are used in a knowledge resource. Four types of interpretant may exist for a given concept in a knowledge resource:

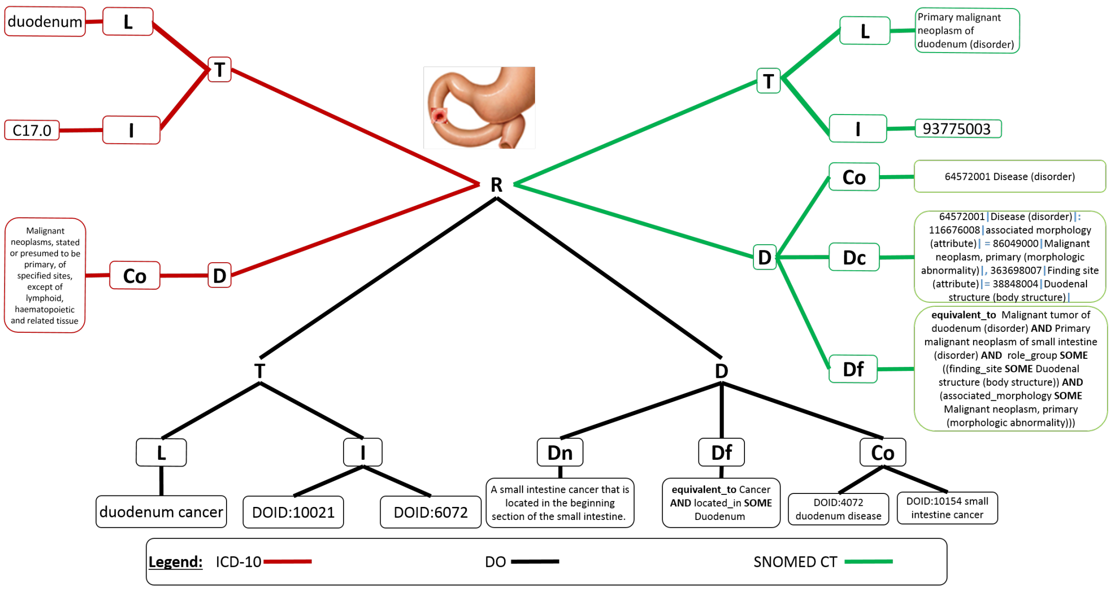
    — Df: the formal description of the concept (given in "Description Logics" [99]). Such a rich description is only available in ontologies.

    — Dn: the description of the concept in natural language (*i.e.*, its textual definition).

[7] For readability, the concepts will be italicized and underlined in this chapter

— Dc: a combination of labels or indexes to express a concept according to the post-coordination mechanism [100, 101]. Some knowledge resources may have a specific computational grammar for defining new concepts with post-coordinated expressions, like in the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [8].

— Co: the context of the concept. It is the hierarchy to which it belongs (the position at one point of the hierarchy). This position is not necessarily given by the structure of the knowledge resource but it may exist implicitly in the guidelines of the resource. It may also be the set of all binary relations in which the concept is involved within the knowledge resource.



duodenum
L
T
C17.0
I
Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, haematopoietic and related tissue
Co
D
R
T
L
Primary malignant neoplasm of duodenum (disorder)
I
93775003
Co
64572001 Disease (disorder)
D
Dc
64572001 | Disease (disorder) | : 116676008 | associated morphology (attribute) | = 86049000 | Malignant neoplasm, primary (morphologic abnormality) | , 363698007 | Finding site (attribute) | = 38848004 | Duodenal structure (body structure) |
Df
equivalent_to Malignant tumor of duodenum (disorder) AND Primary malignant neoplasm of small intestine (disorder) AND role_group SOME ((finding_site SOME Duodenal structure (body structure)) AND (associated_morphology SOME Malignant neoplasm, primary (morphologic abnormality)))
T
L
duodenum cancer
I
DOID:10021
DOID:6072
D
Dn
A small intestine cancer that is located in the beginning section of the small intestine.
Df
equivalent_to Cancer AND located_in SOME Duodenum
Co
DOID:4072 duodenum disease
DOID:10154 small intestine cancer
Legend:
ICD-10
DO
SNOMED CT

Figure 1.2 – Representation of the Malignant neoplasm of duodenum disease according to the <R,T,D> triplet and using its descriptions given in ICD-10, Disease Ontology (DO) and SNOMED CT. The logical definitions (Df) are provided according to the Manchester syntax [102].

1.3.3 Semantic conflicts to be resolved for enriching the integration process

Between two entities represented by the redefined <R,T,D> formalism (<R_1,D_1,T_1> and <R_2,D_2,T_2>), each element of the triplet is a potential level where a semantic

[8] https://confluence.ihtsdotools.org/display/DOCSTART/7.+SNOMED+ CT+Expressions

conflict can appear.

1. The first level concerns the similarity between the representamens (T_1 and T_2). This similarity corresponds to the morphosyntactic similarity of labels (L_1 and L_2). A dissimilarity is established between T_1 and T_2 when a given similarity measure between labels is below a given threshold.
2. The second level relates to the similarity between the interpretants (D_1 and D_2). This similarity corresponds to an identical position of the definition in a semantic hierarchy. This similarity can be calculated by structure-based methods.
3. The last level is the similarity between the two referents (R_1 and R_2) when R_1 and R_2 refer to the same knowledge. Thus, because a referent is the intended concept, this similarity cannot be calculated but it must be interpreted according to the results obtained at the two preceding levels.

From the definition of these similarities and dissimilarities, we can refine the semantic conflicts (Figure 1.3). In particular, two types of conflicts have been added to the naming, scaling and confounding conflicts: the **combined conflicts** (combination of previous conflicts) and **open conflicts** (situations where the concepts to be related are different but complementary).

Absence of dissimilarity.

Two situations can occur when no dissimilarity has been found:

— *Perfect correspondence (R_1=R_2, T_1=T_2, D_1=D_2)*: it is the ideal situation where both triplets contain exactly the same referents (R), representamens (T) and interpretants (D).

— *Combined conflicts between different concepts (only R_1≠R_2)*: it is the situation where the triplets involve concepts that are different. These conflicts can be theoretically found if a difference between concepts has already been established. For example, the parthood[9] relation can lead to combined conflicts [103]. Thus, in the two assertions (derived from the Foundational Model of Anatomy [104]) “skin of hand *part of* hand” and “hand *part of* upper limb”, the notion of *part of* describes a partitive relationship between two elements. However, if the first concept describes a constitutive part relationship, the second describes a regional part relationship [105]. The absence of explicit definitions of these distinct notions can lead to combined conflicts.

[9] http://ontologydesignpatterns.org/cp/owl/partof.owl

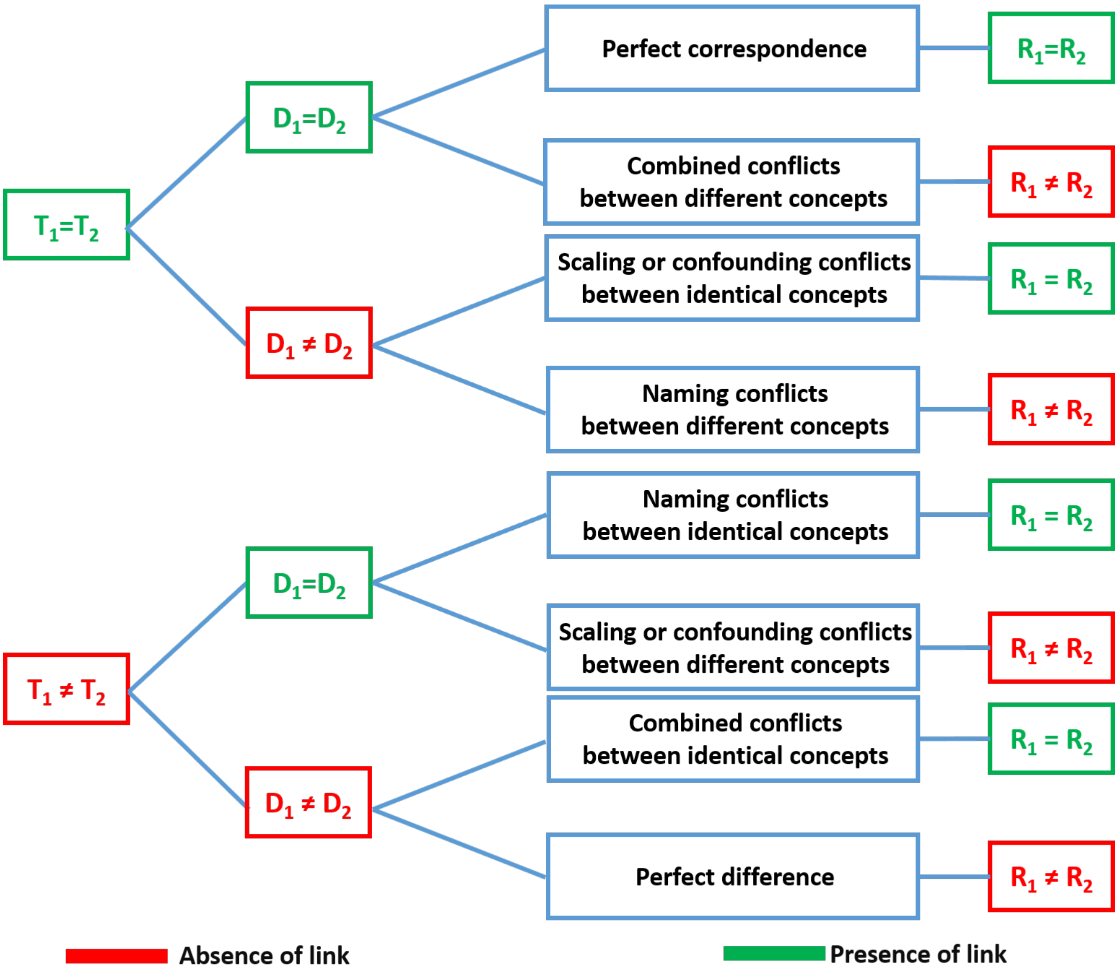

$T_1=T_2$
$D_1=D_2$
$D_1 \neq D_2$
Perfect correspondence
$R_1=R_2$
Combined conflicts between different concepts
$R_1 \neq R_2$
Scaling or confounding conflicts between identical concepts
$R_1 = R_2$
Naming conflicts between different concepts
$R_1 \neq R_2$
$T_1 \neq T_2$
$D_1=D_2$
$D_1 \neq D_2$
Naming conflicts between identical concepts
$R_1 = R_2$
Scaling or confounding conflicts between different concepts
$R_1 \neq R_2$
Combined conflicts between identical concepts
$R_1 = R_2$
Perfect difference
$R_1 \neq R_2$
🟥 Absence of link
🟩 Presence of link

Figure 1.3 – Characterization of semantic conflicts according to the redefined <R,T,D> triplet. Each semantic conflict is described according to similarities or dissimilarities that can occur within each element of the triplet. Absence of dissimilarity induces perfect correspondence (R_1=R_2,T_1=T_2, D_1=D_2) or combined conflicts between different concepts (only $R_1 \neq R_2$). Irrelevant similarity or dissimilarity between representamens results in naming conflicts between different concepts ($R_1 \neq R_2$, T_1=T_2, $D_1 \neq D_2$) and naming conflicts between identical concepts (only $T_1 \neq T_2$). Irrelevant similarity or dissimilarity occurring between interpretants of identical concepts (R_1=R_2, T_1=T_2, $D_1 \neq D_2$) or between different concepts ($R_1 \neq R_2$, $T_1 \neq T_2$, D_1=D_2) induces scaling or confounding conflicts. Absence of similarities results in combined conflicts between identical concepts (R_1=R_2, $T_1 \neq T_2$, $D_1 \neq D_2$) or reveals a perfect difference ($R_1 \neq R_2$, $T_1 \neq T_2$, $D_1 \neq D_2$). In case of perfect difference, this reflects an open conflict when concepts describe complementary notions.

Irrelevant similarity or dissimilarity between representamens

This situation corresponds to naming conflicts in the <R,T,D> formalism.

— *Naming conflicts between different concepts ($R_1 \neq R_2$, $T_1 = T_2$, $D_1 \neq D_2$)*: this situation corresponds to homonymy which occurs when there is a morphosyntactic similarity between labels of different concepts. For example, the term "duodenum" designates the organ in ICD-O3 whereas this term is used in ICD-10 in which it designates the primary malignant neoplasm of this organ (because the ICD-10 code is a subclass of the "Malignant neoplasms" class). Concepts are clearly different although the labels are the same.

— *Naming conflicts between identical concepts (only $T_1 \neq T_2$)*: this often corresponds to *synonymy*. In this case, a conflict appears because two identical concepts have different labels. For example, *Organ heart* is designated by the term "heart" in FMA [10] whereas in SNOMED CT [11], it is the term "cardiac structure", which designates it.

Sometimes, one of the two triplets does not have any label. This case mostly occurs between concepts that are not described in knowledge resources as entities but that can be described by post-coordinated expressions.

For example, the *Diagnosis of benign neoplasm of duodenum* can be designated in ICD-10 by the code D13.2 or by the combination of the codes 8000/0 and C17.0 in ICD-O3. In ICD-10, a label corresponding to this concept exists whereas in ICD-O3, there is no label - even if each element of the post-coordination has a label (8000/0-*Neoplasm, benign* ; C17.0-*Duodenum*).

1.3.4 Irrelevant similarity or dissimilarity occurring between interpretants

These dissimilarities appear when there is a lack or an evolution in the definition of concepts.

— *Conflicts between identical concepts ($R_1 = R_2$, $T_1 = T_2$, $D_1 \neq D_2$)*: when two identical concepts with identical labels present distinct definitions, the situation corresponds to a:

[10] http://xiphoid.biostr.washington.edu/fma/fmabrowser-hierarchy.html?search=none&entryPoint=organSystems&extendHierarchy=true

[11] Look for the SNOMED CT concept 80891009 in:
https://browser.ihtsdotools.org/?perspective=full
&conceptId1=404684003&edition=en-edition&release=v20190131&server=
https://prod-browser-exten.ihtsdotools.org/api/snomed&langRefset=900000000000509007

  - *scaling conflict* if the two definitions are not contradictory but are at different levels of precision. For example, in SNOMED CT, *Ebola disease* is defined as a "Filoviral hemorrhagic fever" and in ICD-10 among "Other viral haemorrhagic fevers, not elsewhere classified". This concept has the same label in the two knowledge resources, being "Ebola virus disease".
  - *confounding conflict* when there is a contradiction in the definitions. For example, *Thalidomide* is represented in ATC [12] among "Antineoplastic and immunomodulating agents", although it has been introduced as an "Hypnotic". Even if the indication of *Thalidomide* has changed, it is still described as a "hypnotic" in SNOMED CT [13].

— *Conflicts between different concepts ($R_1 \neq R_2$, $T_1 \neq T_2$, $D_1 = D_2$)*: scaling or confounding conflicts may arise when two triplets representing two different concepts, with different labels, have identical interpretants. This can be a problem due to the granularity of interpretants (case of scaling conflicts), or this can be a consequence of obsolete or erroneous definitions (case of confounding conflicts). For example, the *Amyotrophic lateral sclerosis* disease is defined in SNOMED CT as a motor neuron disease while the disease *Primary lateral sclerosis* has the same definition in ICD-10. Although these two concepts have the same definition, this is a case of scaling conflict because the concepts are different ("to be a motor neuron disease" is not a sufficient definition for both diseases).

1.3.5 Absence of similarities

Finally, we can describe two situations that can occur when there is no similarity between the representamen and the interpretant ($T_1 \neq T_2$ and $D_1 \neq D_2$):

— *Combined conflicts between identical concepts ($R_1 = R_2$)*: these conflicts correspond to a combination of a naming and a scaling or confounding conflicts. An illustration of such situation is the ICD-10 concept *Malignant neoplasm of Meckel diverticulum*, whose label is "Meckel diverticulum" and which is defined as a malignant neoplasm of the small intestine. The same concept in SNOMED CT is designated by the term "Malignant tumor of Meckel's diverticulum (disorder)" and is defined as a malignant tumor of the ileum, a disorder of the extra-embryonic membrane and a disorder of the embryonic structure. There are no contradictions between the two definitions, but the definition in SNOMED CT is more precise. This conflict thus corresponds to a combination of naming and scaling conflicts.

[12] https://www.whocc.no/atc_ddd_index/?code=L04AX02

[13] September 2018 version and earlier

— *Perfect difference ($R_1 \neq R_2$)*: it arises when two triplets representing two different concepts have different interpretant and labels. This is the "ideal" case where there is no similarity at all between the two <R,T,D> triplets.

This situation is of particular interest when such concepts correspond to complementary knowledge. If no semantic links are defined between them, it will not be possible to relate the concepts, which results in what we chose to call an **open conflict**. For example, it would be useful that the two different concepts *Lyme disease* in ICD-10 and *Genus Borrelia (organism)* in SNOMED CT are linked through a *has_causative_agent* relation.

1.4 Techniques for the creation of mappings

To identify the additional layers needing to be implemented for overcoming the listed semantic conflicts, this section presents existing techniques, used within alignment and integration processes, according to the occurrence of the five semantic conflicts defined in the previous section. Techniques can be grouped into different categories [106, 107]. The most common grouping description of techniques has been proposed by Euzenat *et al.* [107] who distinguish terminological (or lexical), structural, extensional (or instance-based) and semantic techniques. By considering instances like elements of the knowledge resource structure, we address extensional techniques in the same way as structural techniques. Our analysis of the state-of-the-art shows that existing techniques well address naming conflicts but the resolution of the other conflicts remains challenging.

1.4.1 Lexical techniques for establishing morphosyntactic similarity

Lexical techniques [30, 31] generally constitute the first step for establishing mappings automatically. Many strategies have been described in the literature [28, 29, 108] to calculate lexical similarities. These strategies are composed of multiple steps (*e.g.*, lemmatization, tokenization, morphology of syntax, global namespace), each one being useful. However, these strategies depend on the quality of labels. As noted in [109], if reference knowledge resources have stable and well-defined labels, this is not the case for local knowledge resources whose labels may present acronyms and word truncations requiring a pre-process before applying lexical techniques [44].

Lexical techniques are used for the alignment and/or integration of diverse knowledge resources [32, 110]. In the international Ontology Alignment Evaluation Initiative (OAEI), a benchmarking initiative started in 2004, Shvaiko *et al.* have shown by an analysis of the algorithms proposed in the frame of OAEI campaigns that most systems begin the creation of mappings with morphosyntactic approaches [31]. This is the case in SAMBO [111], QODI [112], AgreementMaker [113] and ServOMap [18].

Lexical techniques are however confronted to the following naming conflicts:

— synonymy, which prevents the establishment of correspondences between similar concepts,

— homonymy, which results in correspondences between different concepts,

— concepts generated by post-coordination, which impede the possibility to calculate a similarity measure.

To address the limitations of lexical approaches, some authors combined them with structural techniques [28].

1.4.2 Structural techniques for the resolution of naming conflicts

Structural techniques for dealing with synonymy

Relying on correspondences found by lexical approaches, some authors calculated graph or instance-based proximity between entities [32, 114, 115]. These strategies bring about the possibility to match synonymous entities.

For example, in SAMBO [111], mappings are created between two concepts which lie in a similar position with respect to *is_a* (*i.e.*, subsumption) or *part_of* relationships according to the mappings identified morphosyntactically. ServOMap [18] establishes similarity between concepts having the same structural proximity (*i.e.*, parents, siblings and descendants) according to already mapped concepts. QODI [112] calculates similarity measures according to paths between concepts. Authors select a specific path between two concepts in a first ontology. Then, they compare this path with different possible paths present in the second ontology. The comparison is based on the similarity between: (i) source concepts of each path, (ii) datatype properties, (iii) labels of the concepts located between the source concept and the last concept within each path, and on a penalty for path length difference.

Structural techniques for dealing with homonymy

To reduce errors of lexical techniques, some authors used repair processes [37, 116]. A repair process consists in detecting erroneous similarity between labels for deleting wrong mappings. The main strategy entails highlighting and removing mappings between entities which do not belong to the same context (Co). For example, mappings can be deleted if the entities to be mapped belong to disjoint axes [117] or if they do not comply with predefined reasoning rules [118]. By using algorithms called "reasoners", reasoning consists in inferring an enriched structure of knowledge resources thanks to logical consequences made from explicit assertions.

Structural techniques in case of post-coordination

In addition to the resolution of synonymy and homonymy, structural strategies have been used in the literature to overcome naming conflicts in case of post-coordination [119]. Let us consider an entity e_1 which is decomposed in a given knowledge resource according to its characteristics (*e.g.*, a cancer described according to its morphology and its localization). If each of these characteristics is mapped to an entity of another knowledge resource, a correspondence may be found between e_1 and a combination of the mapped entities of the second resource (corresponding to a post-coordinated expression).

For relating pre-coordinated concepts and post-coordinated expressions, a strategy proposed by Dolin *et al.* consists in providing a canonical form to pre-coordinated concepts and post-coordinated expressions through Health Level 7 Reference Information Model (HL7 RIM) [120]. For the creation of mappings between pre-coordinated concepts and post-coordinated expressions existing in different knowledge resources, Dhombres *et al.* proposed a method that was carried out to increase the coverage of HPO (Human Phenotype Ontology) concepts mapped to SNOMED CT concepts [119]. Authors developed an algorithm for identifying each term that represents a clinical notion within HPO concept labels (*e.g.*, the HPO label "*abnormality of the lip*" contains "*abnormality*" as disorder and "*lip*" as *anatomical structure*). These terms were then mapped to SNOMED CT concepts according to a morphosyntactic method. The resulting SNOMED CT concepts were combined (through post-coordination) to represent some disorders that correspond to the HPO concept (*e.g.*, still for the HPO concept "*abnormality of the lip*", the corresponding post-coordination expression proposed by authors is the following: 64572001-*Disease (disorder)* + 363698007-*Finding site (attribute)* + 48477009-*Lip structure (body structure)*). This method was thus able to find mappings between pre-coordinated concepts from HPO with post-coordinated expressions in SNOMED CT. Note that this method requires the concept labels to be interpretable with a sophisticated syntax and a structure allowing the automatic post-coordination of knowledge resources to be used together.

Limitations of structural techniques

Despite the issues solved by structural techniques, it is important to notice that they require that the knowledge resources to be related have a fairly high-level structure. In addition, results given by these techniques are miscellaneous [28, 121]. Indeed, they depend on lexical techniques which can generate erroneous correspondences (consecutive to naming conflicts, such as homonymy). Secondly, such techniques do not address confounding and scaling conflicts. In fact, the structure of knowledge resources is based on relationships and the interpretants are a set of relations between entities (*e.g.*, formal definitions). Thus, these interpretants may be confronted to quality issues (*e.g.*, erroneous definitions, knowledge evolution) leading to confounding and scaling conflicts.

1.4.3 Semantic techniques: use of lexical and structural techniques in combination with external knowledge

Structural techniques obviously require a minimum of structure within knowledge resources [122, 123]. Thus, overcoming naming conflicts by using structural techniques is limited by the quality of the resource structure. In the literature, some authors harnessed external knowledge to deal with the flaws of the structure of the resources. Using external resources is a good way to detect synonymies [30, 89, 108, 124].

For example, mappings have been validated by multiple experts in a consensual way in [70, 125] and the Health Level 7 Reference Information Model (HL7 RIM) has been used to resolve conflicts due to post-coordination in [120]. In addition, AgreementMaker [113] uses Wordnet [126], a lexical database for English, and UBERON [127], a multispecies anatomy ontology. SAMBO uses also two external resources. The first resource used by SAMBO is the UMLS (Unified Medical Language System) Metathesaurus®, a multi-terminological system containing more than 170 biomedical terminologies, for finding similarities between concepts of knowledge resources that are included in the UMLS. The second resource used by SAMBO is PubMed for calculating similarity measures based on the concept co-occurrence in knowledge resources to be mapped in a set of PubMed abstracts.

Notwithstanding the possibilities offered by the semantic techniques, their use in the literature is generally limited to the identification of identical concepts between different knowledge resources. We believe that they may be useful to find the appropriate relation between complementary concepts.

1.4.4 Remaining challenges to go beyond the integration process

The remaining challenges when trying to relate knowledge resources are the resolutions of scaling, confounding and open conflicts. To describe how such conflicts can be detected, we have drawn inspiration from some works carried out in Web information sharing [79, 128, 129]. The main strategy is based on the structure of ontologies [78, 130]. When reasoning on the ontology structure, some authors detected semantic conflicts through logical errors induced by a mapping process (*e.g.*, mappings between disjoint classes or contradictory properties) [129]. Conversely, other authors made use of unsatisfiable concepts for identifying erroneous placements of entities within the hierarchy thanks to reasoning [79, 129].

Resolution of scaling and confounding conflicts

As described within Web information sharing, ontologies provide knowledge on unstructured data allowing their automatic interpretation [131]. Therefore, the knowledge supplied by a support ontology can be described as a formal definition for each entity to be linked. To resolve scaling and confounding conflicts, it is possible to formalize conflict resolution strategies induced by the formal definition of entities [88]. These strategies can be resumed by the interpretation and transformation of the interpretants according to the logical structure of the support ontology. Thus, scaling or confounding conflicts can be detected by confronting the interpretant of the entities to be mapped to those in the support ontology.

Then, these conflicts can be resolved by providing logical definitions or additional definitional features[14] coming from the support ontology to the entities to be mapped (Figure 1.4). This strategy obviously assumes that the support ontology covers the domains of the knowledge resources to be integrated and also provides appropriate granularity and quality for the interpretants of its entities.

[14] Definitional features correspond to the hierarchical context (Co) of entities used to describe another entity. For example, in the National Cancer Institute thesaurus (NCI thesaurus), "anatomical localization" and "histological lesion" are definitional features that can be used for describing "cancer disease" entities

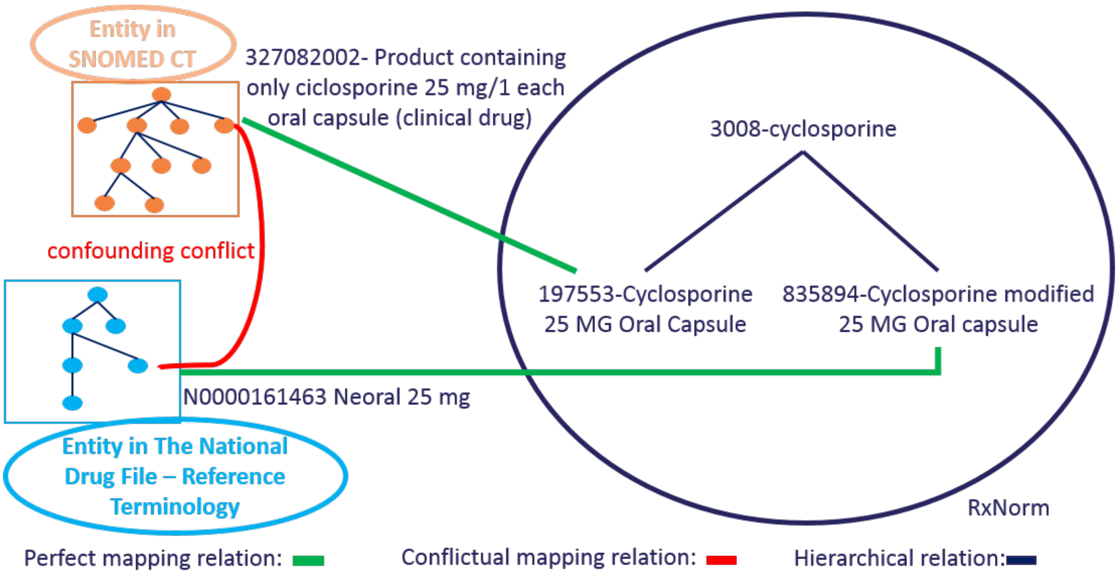

Entity in SNOMED CT
327082002- Product containing only ciclosporine 25 mg/1 each oral capsule (clinical drug)
3008-cyclosporine
confounding conflict
197553-Cyclosporine 25 MG Oral Capsule
835894-Cyclosporine modified 25 MG Oral capsule
N0000161463 Neoral 25 mg
Entity in The National Drug File – Reference Terminology
RxNorm
Perfect mapping relation: 🟩
Conflictual mapping relation: 🟥
Hierarchical relation: ▬

Figure 1.4 – Illustration of a confounding conflict highlighted by RxNorm as a support knowledge resource. In medical practice, the use of “cyclosporine modified” in spite of (non-modified) cyclosporine is clinically different and must be specified. Neoral being the brand name of a clinical drug containing “cyclosporine modified”, the equivalence between the entity in SNOMED CT and the entity in NDF-RT is erroneous.

Resolution of open conflicts

In cases of open conflicts, to find links between complementary concepts already known as different, some authors used external resources, such as a model and/or a support ontology [35, 36]. Thus, for linking different but complementary entities which correspond to elements of the model or entities of the support ontology, transversal relations can be used (Figure 1.5).

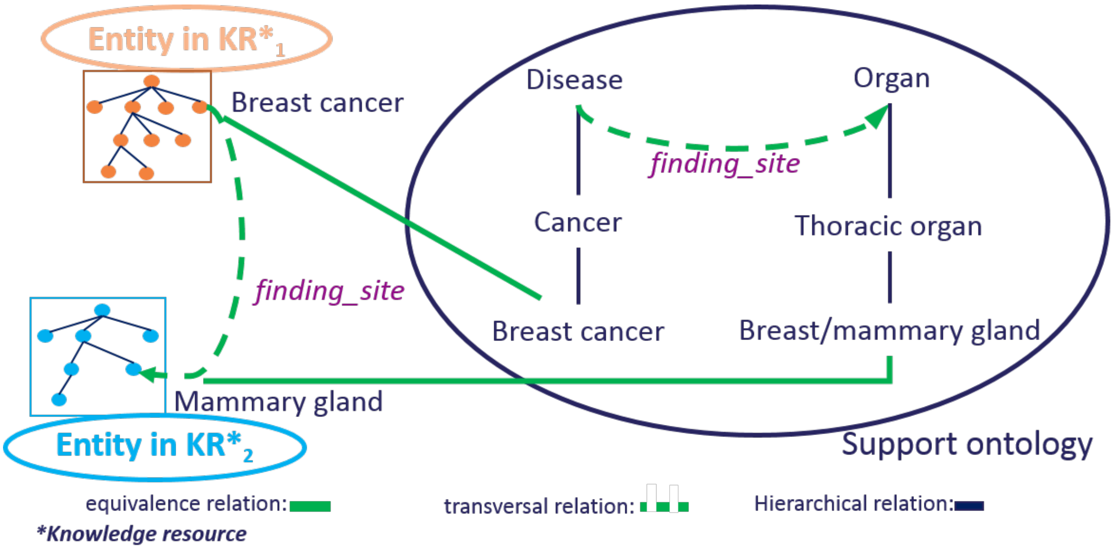


Figure 1.5 – Illustration of the resolution of an open conflict using a support ontology. The transversal relation *finding_site* is used to relate the main entities of “Breast cancer” and “Mammary gland”, which are different but correlated. It is important to notice that the naming conflict between “Mammary gland” and “Breast” is resolved by the support ontology which use both terms to designate the concept. KR means “knowledge resource”.

1.5 Conclusion

In this chapter, two main aspects have been explored: (1) the formal framework for the description of the content of knowledge resources, and (2) the different processes that can be performed to overcome semantic conflicts between knowledge resources.

We used the semiotic triad of Peirce [95] to describe the concepts of knowledge resources. Representing the contents of ontologies [132] and/or terminologies [46] with semantic triplets has already been realized in the literature. In Visser *et al.* [132], the semantic triplet is specific to ontologies whereas the triplet that we propose concerns all types of knowledge resources. It can thus be seen as an abstraction of the triplet proposed by Visser *et al.*. In addition, the referents in this representation correspond to entities in the ontologies and not to the referred concepts of each of them, which is the case in our work. Using the referred concepts in the related domain as the referents gives the possibility to represent all the concepts induced by the structure of knowledge resources, in addition to the knowledge represented by each entity. Considering the two separated dimensions of terminologies (linguistic dimension) and ontologies (conceptual dimension), Roche [46] introduced the notion of "Ontoterminology" using a double semantic triangle. In our work, the use of Pierce's semiotic triad helps reconciling the two dimensions in the common framework <R,T,D>, in which each dimension conserves its characteristics. Our approach responds, above all, to a practical need for the joint use of knowledge resources, regardless of their philosophy of creation.

Based on this formalism and through an analysis of the literature describing existing processes for finding correspondences between knowledge resources regardless of their structure, we refined the description of semantic conflicts and illustrated the necessity for semantically enriching the integration process. The objective of such enhancement is to identify appropriate relations to link different but complementary concepts. Support ontologies appear as the best solution to deal with all the requirements of a process that can overcome any type of semantic conflicts.

Considering this assumption, we present in the next chapters the use of a knowledge resource as a support for each identified process: the alignment (chapter 2), the integration (chapter 3) and the enriched integration (chapter 4). We intend to highlight how each process deals with semantic conflicts according to its objective.

I know what I have given you... I do not know what you have received

Antonio Porchia

Chapter 2

Alignment process: application to the biological analyses

Summary: In this first implementation, we aligned the interface terminology of the Bordeaux university hospital to Logical Observation Identifiers Names and Codes (LOINC). This work describes the overcoming of noisy labels available in the interface terminology for the alignment process.

We firstly constructed a graph structure for LOINC using its standardized labels. This stage consisted in automatically incorporating the naming rules of LOINC labels, based on punctuation. We implemented these rules and applied them on French versions of LOINC.

Secondly, we pre-processed the noisy labels of the interface terminology. This stage consisted in applying strategies developed for processing texts in forums, social networks and short message systems to non-standard words that are used in the interface terminology of the Bordeaux university hospital. Thus, the main aspect in this part concerned the resolution of naming conflicts to find the appropriate mappings according to equivalence or subsumption relations between the entities to be related.

Finally, we used the constructed graph structure of LOINC and the enhanced labels of the interface terminology in an alignment process based on a morphosyntactic mapping step using the ServoMap tool and a filtering step based on the LOINC structure.

Keywords interface terminologies, non-standard words, LOINC, alignment process

Valorization The work described in this chapter has been valued in the frame of two articles. The first article, entitled “Processus de prétraitement des libellés d'une terminologie d'interface", has been published in the proceedings of the *Symposium sur l'Ingénierie de l'Information Médicale* (SIIM), held in Toulouse in 2017. The last article, entitled "Alignment of an interface terminology to the Logical Observation Identifiers Names and Codes (LOINC®)" is currently in preparation for a submission to the Journal of the American Medical Informatics Association.

2.1 Introduction

In this chapter, we present our first implementation that concerned the alignment of the Logical Observation Identifiers Names and Codes (LOINC®) to the "interface terminology" used to encode biological analyses in the university hospital of Bordeaux. Thus, we describe a strategy used to overcome the naming conflicts between "reference terminologies" and interface terminologies. This strategy consisted in enhancing the quality of labels, applying morphosyntactic techniques to find mappings and finally using structural techniques based on LOINC's structure to perform a repair process.

Interface terminologies are controlled vocabularies, which have been defined in the biomedical domain as follows: *"a systematic collection of health care-related phrases (terms) that supports clinicians' entry of patient-related information into computer programs"* [133, 134]. Indeed, such terminologies are created for specific use cases within some given healthcare structures. If the usability of interface terminologies is important for the health information systems in which they are developed, their use may be limited in an integrated perspective. For interoperability purpose, interface terminologies have to be aligned to **reference terminologies** [133, 135], which are consensual knowledge resources whose terms and structures have been validated by the scientific community. Thus, aligning an interface terminology to a reference terminology is required for sharing data between different health information systems [136, 137]. In the literature, many works have been concerned with this issue [134, 138].

The ideal way to get an interface terminology aligned to a reference terminology is to directly create the interface terminology from a reference terminology [139–141]. But most of the time, this strategy cannot be applied. Indeed, interface terminologies are usually created manually using items present in paper forms [136]. Consequently, it is necessary to adapt techniques commonly used in the literature for finding correspondences between terminologies [31], to align interface terminologies to reference terminologies. At the Bordeaux university hospital, such an interface terminology is used for encoding and retrieving results of biomedical analyses. This interface terminology is herein referred to by its French acronym TLAB for "Terminologie Locale d'Analyses

Biomédicales".

Many characteristics can induce the selection of a reference terminology as a support for sharing information. Some reference terminologies are created and/or recommended by the World Health Organization (WHO), such as the ICD-10 [1], which is used worldwide for epidemiology purpose. Nevertheless, the novelty and the quality of some terminologies have imposed themselves as a reference in their sub-domain. LOINC is an example of such knowledge resources for recording laboratory observations in many countries [14, 15]. Containing validated terms of the domain, LOINC is a reference terminology. Thus, many works have been concerned by the mapping of local terminologies to LOINC [142–145], positioning LOINC as an international support knowledge resource for sharing information across different health systems. The selection of LOINC within our alignment process has consequently been motivated by its wide-scale adoption and use for representing biological analyses in a standardized way.

In the next section, the characteristics of LOINC and TLAB are presented, as well as existing approaches for aligning terminologies in the light of these characteristics. Then, we present the materials that we used and the methods that we developed for the alignment process in sections 2.3 and 2.4. Finally, we present the main results we obtained before concluding this chapter.

2.2 Background

This section describes the characteristics of TLAB and LOINC and discusses existing techniques that have been developed for aligning an interface terminology to a reference terminology.

2.2.1 Terminologies to be aligned

TLAB

The interface terminology used at the Bordeaux university hospital for encoding data of the medical test laboratory has been exported from the electronic health record system of the hospital. TLAB labels are described in French and have been recorded manually by health professionals. The space limits in the recording step lead to non-conventional abbreviations of labels (*e.g.*, *PCR.C.TRACHO/GENI*).

TLAB is a multi-axial terminology composed of 29,227 entities that are hierarchically organized. The absence of formal descriptions for TLAB entities

[1] https://icd.who.int/browse10/2016/en#/

makes the Simple Knowledge Organization System (SKOS [2]) format adequate to represent TLAB [146]. Thus, TLAB entities have been described as *skos:Concept* and their hierarchical relations have been defined through the *skos:broader* relationship. Each entity corresponds to an alphanumeric code (Index) (*e.g.*, synana-vrpu1). Among them, 29,191 indexes (I) are related to a label (L) using the *skos:prefLabel* attribute, which means that 36 indexes (I) have no associated label.

Only 8,285 entities of TLAB are rooted by one of the following 15 high-level entities that correspond to the different domains of biological analyses:

1. *Anatomie et Cytologie Pathologiques (Pathological Anatomy and Cytology),*
2. *Bactériologie (Bacteriology),*
3. *Biochimie (Biochemistry),*
4. *Immuno-hématologie EFS (Immunohematology),*
5. *Génétique (Genetic),*
6. *Hématologie (Hematology),*
7. *Immunologie - Immunogénétique (Immunology and Immunogenetics),*
8. *Mycologie - Parasitologie (Mycology - Parasitology),*
9. *Hormonologie - Marqueurs tumoraux (Hormonology - Tumor markers),*
10. *Biologie de la reproduction (Reproductive biology),*
11. *Pharmacologie - Toxicologie (Pharmacology - Toxicology),*
12. *Recherche (Research),*
13. *Biologie des tumeurs (Tumor biology),*
14. *Virologie (Virology),*
15. *Hygiène hospitalière (Hospital hygiene).*

When the extraction process could not associate an entity to one of these 15 high-level entities, it was described as being an orphan. The hierarchical structure of TLAB being important in the process of alignment, these unclassified entities were excluded from the alignment process described in section 2.4.

Thus, the 8,300 TLAB entities considered in the alignment process have the following characteristics:

- 5 indexes have no label,
- 7,202 distinct labels exist, of which 639 are related to several indexes with a maximum of 59 indexes for a given label.

[2] https://www.w3.org/TR/skos-reference/

LOINC®

LOINC is a reference terminology created and maintained by the Regenstrief Institute [15]. Published in 1995 [147], the first release of LOINC contained only codes for laboratory testing. Nowadays, LOINC is a clinical terminology for recording health measurements, observations and documents [15]. The current version contains 50,000 codes describing lab tests for a total of 89,271 codes. That codes are hereafter designated as “LOINC concepts”. The LOINC concepts are defined using the following attributes (Figure 2.1):

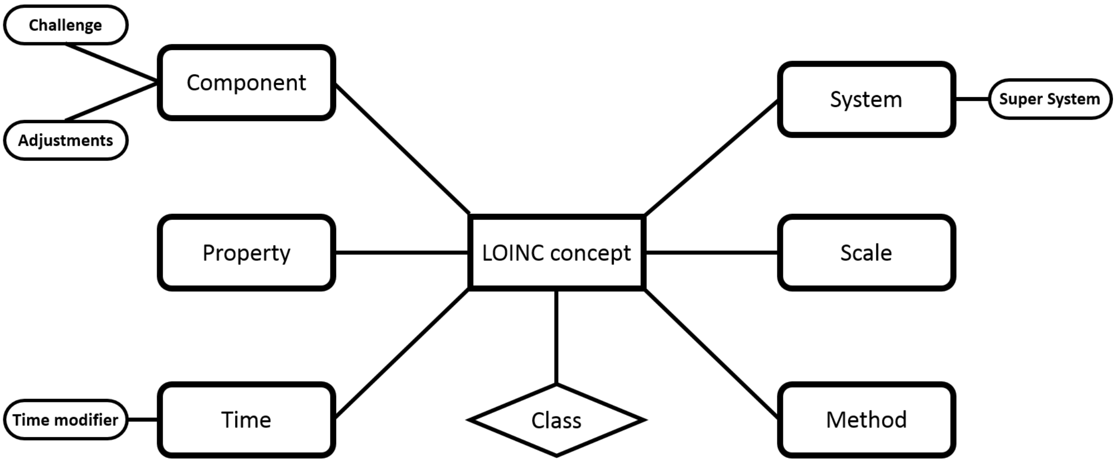


Figure 2.1 – The description model of LOINC concepts. The model contains six mandatory attributes (rectangles with rounded corners) and four optional attributes (ovals) to refine the description of three mandatory attributes (component, system and time). Each LOINC concept is attached to a specific class. In the LOINC users’ guide [148], it is indicated that classes are not definitional for LOINC concepts but that they are used for sorting purpose. These classes mainly correspond to the analysis type of a lab test.

— six major attributes:

1. Component: it corresponds to the measured or observed analyte (*e.g.*, sodium, ABO group, creatinine renal clearance),
2. Property: it represents the different quantitative and qualitative measurements (*e.g.*, mass, entitic number, catalytic activity, entitic volume),
3. Time: it indicates the punctual or the interval characteristics of the measurement (*e.g.*, point in time, episode, less than 1 hour, 6 hours),

4. System: it mainly corresponds to the sample or the body system (*e.g.*, abscess, blood venous, eye),
5. Scale: it provides a precision of the observation of the measurement (*e.g.*, quantitative, ordinal, narrative, nominal),
6. Method: it corresponds to the technique applied to obtain the results, if it is clinically relevant to notice it (*e.g.*, agglutination, immune fluorescence, visual count).

— four minor attributes:

1. Challenge: it describes the possible preliminary action to realize before the test.
2. Adjustments: it is used when some specific corrections are realized on the obtained values.
3. Time modifier: the values taken by this optional attribute are *min*, *max*, *first*, *last* and *mean* (default value).
4. Super-system: it specifies the origin of the sample, if it is not a patient.

LOINC concepts corresponded to a unique identifier (Index I) with a fully specified name (label L), corresponding to the concatenation of LOINC attributes according to a specific order and to specific punctuations (*e.g.*, ":" to separate major attributes as follows: <Analyte/component>:<kind of property of observation or measurement>:<time aspect>:<system (sample)>:<scale>:<method> and "$\wedge$" to describe a minor attribute).

The LOINC concepts for lab tests can be identified using the following 14 classes [14]:

1. *Antibiotic susceptibilities,*
2. *Blood bank,*
3. *Chemistry,*
4. *Coagulation,*
5. *Cytology,*
6. *Fertility,*
7. *Flow cytometry cell markers,*
8. *Hematology cell count,*
9. *Microbiology,*
10. *Molecular pathology,*
11. *Skin tests,*
12. *Pathology,*

13. *Drug & toxicology,*
14. *Urinalysis.*

The labels of LOINC concepts were originally available in English. In the literature, LOINC is described as using a "part-based translation principle" to automatically generate the fully specified name of LOINC concepts. "*The atomic elements that make up each LOINC term name are called Parts*" [15]. LOINC parts mainly correspond to LOINC attributes to which an identifier has been assigned. For example, in the fully specified name of the LOINC concept 3665-7-*Gentamicin^trough:MCnc:Pt:Ser/Plas:Qn*, the atomic elements "*Gentamicin*" and "*trough*" are LOINC parts that are respectively identified by the code LP15747-6 and LP20176-1. Like *Gentamicin* and *trough*, all the delimited elements of the label (*i.e.*, *MCnc*, *Pt*, *Ser*, *Plas* and *Qn*) are LOINC parts. Thus, LOINC editors recommend to firstly translate LOINC parts and then to use these translations for reconstituting the full label of LOINC concepts [149].

2.2.2 Alignment strategies

Relating interface terminologies to reference terminologies is an important and time-consuming task. Automatic strategies are required for this task [142]. The strategy that we had to implement for the alignment of TLAB and LOINC had necessarily to deal with the differences of their structure and the absence of overlap between terms available in the reference and interface terminologies, as well as the lack of quality that may exist in labels of interface terminologies [44, 109].

Existing alignment approaches to LOINC

Many works have been described in the literature using LOINC as the reference terminology for the mapping of laboratory terms [144, 150–154]. Three main strategies are generally used to perform these alignments:

— the manual mapping of interface terminologies to LOINC [150], which is a tedious task, which is not reasonable to implement when dealing with large interface terminologies.

— the use of the REgenstrief LOINC Mapping Assistant (RELMA) [144, 152, 154]. RELMA is an open access mapping tool provided by the Regenstrief Institute for the alignment of local terms (*i.e.*, terms available in interface terminologies or in corpora) to LOINC concepts [145]. RELMA uses a morphosyntactic strategy with a manual correction of mappings, thus needing users' intervention [142]. In practice, the tool firstly proposes LOINC concepts as potential equivalences for local labels (one at a time).

Then, a validation is asked to users or an alternative label entry when no LOINC concept is proposed.

- the use of home-made algorithms [142, 151, 153]. As RELMA, the other mapping strategies are based on morphosyntactic approaches, sometimes combined with machine learning algorithms. Existing morphosyntactic approaches were however ineffective to deal with noisy labels. Indeed, authors that used home-made algorithms and/or RELMA reported that the variation of local terms and the incompleteness of the description in interface terminologies are the main issues altering the quality of mappings. To compensate for these limitations, some of these authors cleaned and enhanced manually the terms in interface terminologies [145, 152].

All the applied strategies were designed rather to increase the number of obtained mappings than to obtain an optimal semantic quality of resulting mappings. Thus, naming conflicts were not overcome by existing automatic processes. We believe that using the structure of LOINC labels as an element of the validation of mappings' may be a solution to deal with this issue. The goal of our process is thus to implement a specific and automatic process for the alignment of TLAB using a corrective step of TLAB labels that takes into account the structure of LOINC for the validation of mappings.

Strategies for the pre-processing of the labels in interface terminologies

The morphosyntactic approach is the common initial step of all automatic mapping processes. Such approaches are limited for interface terminologies because of the quality of their labels [109]. A pre-processing is thus necessary to improve the efficiency of mapping strategies and to overcome naming conflicts. For interface terminologies, it is sometimes possible to find guidelines describing the naming conventions of their labels [44]. If such guidelines are not available (which is the case of TLAB), strategies developed for processing texts in forums, social networks and Short Message Systems (SMS) can be used to improve the quality of local labels [155]. They consist in the detection and correction of "Non Standard Words" (NSWs) [156]. Gadde *et al.* [157] explained the occurrence of NSWs by four situations: character deletion (*e.g.*, "Meningo" for "meningococcus"), phonetic substitution (*e.g.*, "2morrow" for "tomorrow"), abbreviations (*e.g.*, "HIV" for "human immunodeficiency virus") and dialectical usage (*e.g.*, "gonna" for "going to").

2.3 Materials

2.3.1 LOINC release

For our work, we used the LOINC_2.65 version [3], which contained:

— the LOINC core table (LCT) which was the full version of LOINC describing labels in English,

— the French variant tables (FVT) which described the main core of LOINC in French,

— the LOINC part table (LPT) which described the relation of LOINC concepts with their attributes (called "LOINC Parts" in the release).

The release contained four French language variants (French, Belgian, Canadian and Swiss variants). Like in LCT, each line of FVT was composed of a LOINC concept identifier and the labels of the six main attributes (each attribute and the LOINC concept identifier being separated in distinct columns). The French variant contained 49,437 LOINC concepts, the Belgian variant 45,779 LOINC concepts, the Canadian variant 45,411 LOINC concepts and the Swiss variant 4,940 LOINC concepts. The Swiss variant has not been used in our process because only short names (*i.e.*, labels without the punctuation structure) of LOINC concepts was provided. By pooling the French, Belgian and Canadian variants, the French version of LOINC finally contained 54,480 LOINC concepts.

LPT described LOINC concept identifiers and labels, the identifiers and labels of their related LOINC attributes, as well as the type of link existing between LOINC concepts and their attributes (*e.g.*, *component* when the relation holds between a LOINC concept and a component attribute). LPT has already been used to describe a formal structure of LOINC like in Bioloinc [4] and BioPortal [5]. In LPT, a LOINC concept can be related to multiple LOINC attributes of the same type. The additional tags "primary" and "search" further specify the link between LOINC concepts and attributes. The explanatory note accompanying the LPT gives the following two definitions:

— "***Primary***-*the primary parts associated with a given LOINC term, including the six major parts*",

— "***Search***-*parts that are only linked to a given term in order to facilitate efficient searching and location of that term*".

[3] https://loinc.org/

[4] https://bioloinc.fr/bioloinc/KB/#Group:uri=http://aphp.fr/Bioloinc/JDV_LOINC_Biologie;tab=props

[5] https://bioportal.bioontology.org/ontologies/LOINC

Thus, the "search" tag is mainly navigational whereas the "primary" tag is definitional. Note that this does not exclude that a "primary" tag can be found between a LOINC concept and more than one LOINC attribute of the same type.

2.3.2 ServoMap

ServoMap is a mapping tool developed in our research team by Diallo et *al.* [18]. It is *a highly configurable large scale ontology matching system able to process large knowledge resources associated with multilingual terminologies.* ServoMap is based on Lucene [158] and provides equivalence mappings between entities of two terminologies. ServoMap firstly measures morphosyntactic similarity to find a first set of equivalences and then computes structural proximity between entities identified as equivalent. We used the 2013 version of ServoMap in our alignment process.

2.4 Alignment methods

The developed methods are composed of three main stages: (1) the construction of the French structure of LOINC, (2) the pre-process of TLAB labels, and (3) the alignment process.

2.4.1 Construction of the French structure of LOINC

The construction was performed according to the rules based on the punctuation present in the labels of LOINC concepts. We chose to describe LOINC in the SKOS format to be able to represent multiple labels (L) for a given index (I). In addition, the generalization hierarchy in SKOS seemed to be appropriate for representing the LOINC hierarchy that is based on the compositional structure of its labels. Indeed, SKOS has been proven to be well adapted for the description of the hierarchical structure of terminologies [159], and LOINC has already been described using SKOS [160].

We followed two steps to construct the French structure of LOINC: (i) the description of the LOINC model, and (ii) the instantiation of the model according to the structure of LOINC labels.

Description of the proposed LOINC model

Figure 2.2 displays the proposed model for the description of LOINC concepts.

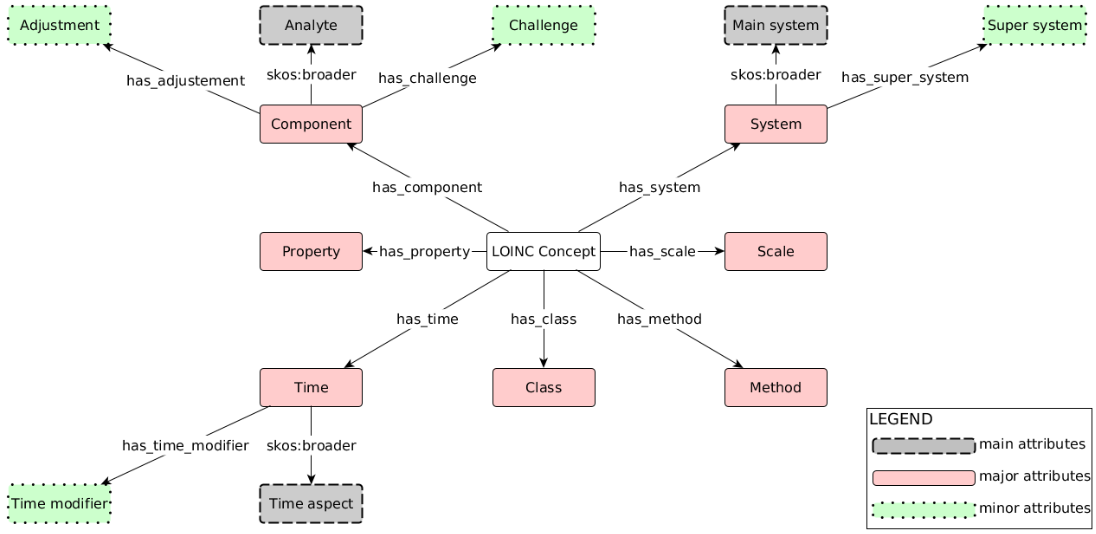


Figure 2.2 – Proposed model used for the construction of the LOINC graph structure

- **The description of LOINC attributes**: in the literature, LOINC attributes are described as ***major attributes*** or ***minor attributes***. When used, minor attributes are parts of the description of major attributes. Thus, they correspond to optional sub-parts of major attributes. We designate hereafter as ***sub-attributes*** the sub-parts of LOINC major attributes. The sub-attributes used for the description of major attributes that are not minor attributes are called ***main attributes*** in the rest of this chapter. For example, in the component (major attribute) "leukocytes^^corrected for nucleated erythrocytes", "corrected for nucleated erythrocytes" is an adjustment (*i.e.*, a minor attribute), while the other part of this component (*i.e.*, "leukocytes") is the main attribute.

- **The description of relations in the model**: for each major attribute, we created a semantic link between the LOINC concept and the attribute. The relation has been labelled using the prefix *"has_"* followed by the name of the major attribute (*e.g.*, a *has_method* relationship has been defined to associate LOINC concepts to their method attribute(s)). Between major attributes and minor attributes, the same strategy has been used (*e.g.*, a *has_adjustment* relationship has been defined to associate component attributes to their related adjustment attribute). Between major attributes and their main attribute, a hierarchical relation has been created (skos:broader).

All LOINC attributes and concepts were described as *skos:Concept*.

Instantiating the model

The instantiation was applied using data from the FVT. The process followed the following two steps:

- **The creation of sets of labels**: for each LOINC concepts, we created sets of labels from all the French variants of each type of attributes using a tokenization process based on the punctuation of LOINC labels. The sets of labels corresponding to main attributes were included in sets of labels of major attributes. For example, we created a unique set of labels for components and analytes.
  - the caret character "^" delimits the minor attributes and the main attributes in the description of major attributes. Using this punctuation, we created the set of major attributes and the set of minor attributes. For example, from the LOINC label of the component *leukocytes^^corrected for nucleated erythrocytes*, "leukocytes^^corrected for nucleated erythrocytes" and "leukocytes" were integrated in the set of components and "corrected for nucleated erythrocytes" was integrated in the set of adjustments.
  - the dot character "." describes hierarchical relations between attributes. For each dot character in a label, a sub-attribute corresponding to the left side of the dot character was created. For example, from the component label "epithelial cells.renal", we created the label "epithelial cells" and includes both labels in the set of components.
  - the slash character "/" describes quotient relations in the components. Like for the dot character, the left side of the slash character was also extracted.
  - the "+" and "&" signs may be used to create combined measurements or combined results. The string characters related by the "+" and/or "&" signs can then be decomposed. The left side of the related characters is identified as a prefix (an identifier was created for that prefix) and the right side as a suffix. The sub-attributes were thus reconstituted by combining the prefix, each related character and the suffix. For example, from the component label *human papilloma virus 16+18 Ag*, the related characters are "16" and "18", the prefix is "human papilloma virus" and the suffix is "Ag". The following labels were thus created and included in the set of components (in addition to the suffix and component label): "human papilloma virus 16 Ag" and "human papilloma virus 18 Ag".

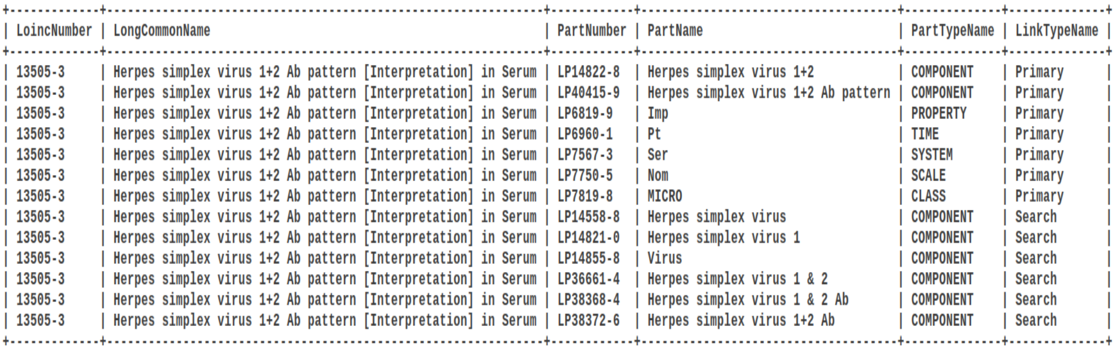
  - — the "+" and "-" signs may be used to describe the cluster of differentiation (CD) of cells when they appear at the end of a label. In such cases, the "+" sign indicates the presence of a specific CD and "-" indicates its absence. Thus, we applied the same rule as for the "+" and "&" signs to identify the composed sub-attributes. For example, from the component label *Cells.CD3+CD4+CD27-CD45RO+CD62L-*, we created the following labels and included them in the set of component: "cells", "cells cd3", "cells cd4", "cells cd27-", "cells cd45ro" and "cells cd62l-".

- — **The creation of an identifier**: For each LOINC concepts and in each set of labels, a unique code was assigned to each label. For each LOINC concept related to an attribute, this attribute was considered as equivalent across the different linguistic variants. Then, a unique identifier was created for this attribute.

  For example, in the FVT, the term "hémostase" was used in the French variant to designate the class attribute of the LOINC concept 3245-8-*Clot Retraction [Time] in Blood by Coagulation assay* while the term "coagulation" was used in the Canadian variant. Thus, a unique code (CLAS1508) was created for the two labels "hémostase" and "coagulation".

Comparison of the constructed structure with the stated structure of LOINC

To highlight the advantages of the constructed structure, we compared it to the stated structure of LOINC, which is commonly computed from the LPT. Each row in the LPT describes a LOINC concept and its related attributes (Figure 2.3).



```
+-------------+---------------------------------------------------------------+------------+-------------------------------------+--------------+--------------+
| LoincNumber | LongCommonName                                                | PartNumber | PartName                            | PartTypeName | LinkTypeName |
+-------------+---------------------------------------------------------------+------------+-------------------------------------+--------------+--------------+
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP14822-8  | Herpes simplex virus 1+2            | COMPONENT    | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP40415-9  | Herpes simplex virus 1+2 Ab pattern | COMPONENT    | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP6819-9   | Imp                                 | PROPERTY     | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP6960-1   | Pt                                  | TIME         | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP7567-3   | Ser                                 | SYSTEM       | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP7750-5   | Nom                                 | SCALE        | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP7819-8   | MICRO                               | CLASS        | Primary      |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP14558-8  | Herpes simplex virus                | COMPONENT    | Search       |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP14821-0  | Herpes simplex virus 1              | COMPONENT    | Search       |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP14855-8  | Virus                               | COMPONENT    | Search       |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP36661-4  | Herpes simplex virus 1 & 2          | COMPONENT    | Search       |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP38368-4  | Herpes simplex virus 1 & 2 Ab       | COMPONENT    | Search       |
| 13505-3     | Herpes simplex virus 1+2 Ab pattern [Interpretation] in Serum | LP38372-6  | Herpes simplex virus 1+2 Ab         | COMPONENT    | Search       |
+-------------+---------------------------------------------------------------+------------+-------------------------------------+--------------+--------------+
```

Figure 2.3 – Description of the LOINC concept 13505-3-*Herpes simplex virus 1+2 Ab pattern [Interpretation] in serum* in the LPT. Screen-shot of the SQL (Structured Query Language) query : "*Select * from LPT where LoincNumber="13505-3"*

For the construction of the LOINC structure, we described each relation between LOINC concepts and attributes as a simple Resource Description Framework (RDF [6]) triple. For each type of attribute, we integrated in the structure only relations between LOINC concepts related to exactly one LOINC attribute through the link tag "primary".

For example, the LOINC concept 13505-3-*Herpes simplex virus 1+2 Ab pattern [interpretation] in serum* has been linked to the class LP7819-8-*Micro*. However, this LOINC concept has not been related to the two components LP14822-8-*Herpes simplex virus 1+2* and LP40415-9-*Herpes simplex virus 1+2 Ab pattern* because they are both tagged as "primary".

The attributes created by our process and those from the LPT were considered as equivalent when they shared the same LOINC concept identifier. Thus, we computed the cardinality of these mappings where:

— 1-0 mappings corresponded to one created attribute for which no stated attribute existed,

— 1-1 mappings associated one created attribute to one stated attribute,

— 1-N mappings associated one created attribute to more than one stated attribute.

2.4.2 Pre-processing of TLAB labels

For the pre-processing of TLAB labels, we had to deal with three of the four situations of NSWs listed by Gadde *et al.* [157] occurring in TLAB labels, being the character deletion (*e.g.*, "Conc.Nucl.In.Plaq"), abbreviations (*e.g.*, "ARN VIH LCR") and the dialectical usage (*e.g.*, "Cyto&CultureUrinAst"). The strategy for correcting and enhancing the quality of labels in TLAB was composed of three stages: (i) the detection of NSWs, (ii) the proposition of corrections for each NSW, and (iii) the correction of NSWs in each label. The objective was to obtain more expressive labels (*e.g.*, to obtain the label (or a close one) "*Recherche par Réaction en chaîne par polymérase de Neisseria gonorrhoeae au niveau génital*" from the label "*R.pcr.N.gonoGeni*").

Detection of NSWs

The detection strategy was based on three sub-processes (Figure 2.4):

— The selection of **potential NSWs**: potential NSWs correspond to words (*i.e.* to strings delimited by white-spaces) that are composed of consonants only or interspersed by punctuation or uppercase letters only.

[6] https://www.w3.org/TR/rdf-schema/

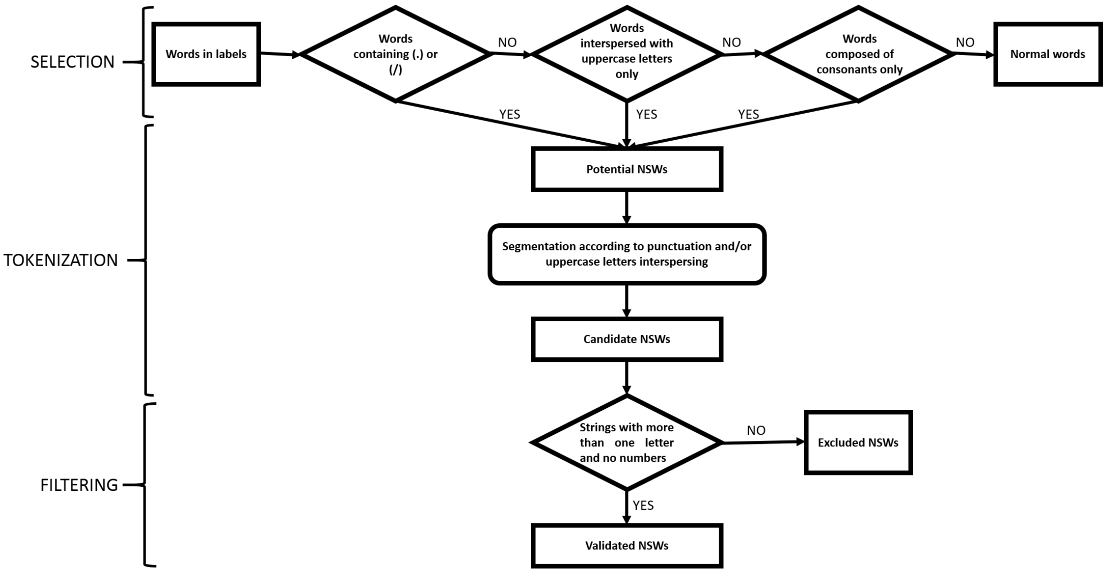

SELECTION
Words in labels
Words containing (.) or (/)
NO
Words interspersed with uppercase letters only
NO
Words composed of consonants only
NO
Normal words
YES
YES
YES
Potential NSWs
TOKENIZATION
Segmentation according to punctuation and/or uppercase letters interspersing
Candidate NSWs
FILTERING
Strings with more than one letter and no numbers
NO
Excluded NSWs
YES
Validated NSWs

Figure 2.4 – The detection process of NSWs. A word corresponds to a string delimited by white-spaces. Potential NSWs are words that are composed of consonants or interspersed by punctuation or uppercase letters only. Candidate NSWs are tokens obtained after segmentation of potential NSWs. Validated NSWs are candidate NSWs that are composed of more than one character or do not contain numbers.

- The tokenization: **candidate NSWs** corresponded to tokens that result from the segmentation of potential NSWs according to interspersing punctuation and uppercase letters.
- The filtering: **validated NSWs** correspond to candidate NSWs that do not contain numbers or consist in at least two characters.

Correction of NSWs

The correction strategy consisted in normalizing the NSWs and in creating a dictionary with these normalized NSWs as entry indexes. The normalization consisted in lowercasing tokens and in suppressing the punctuation attached to each token. The dictionary was created using the list of French medical abbreviations and their corrections given by Wikipedia [7]. In addition, a manual correction was proposed for indexes that are linked to more than five labels in order to enrich the dictionary.

Correction for labels

The corrections of abbreviations found in Wikipedia were systematically used to replace in the labels all corresponding words, described as an abbreviation in the Wikipedia list. This replacement was performed even if the word was not identified as a NSW. For the identified NSWs, the corresponding index was used to obtain their appropriate correction in the dictionary. Each correction lead to an additional label that was integrated using the *skos:altLabel* attribute.

2.4.3 Alignment process

To realize the alignment of TLAB and LOINC, the following three steps further detailed below have been performed:

1. The mapping of tokens constituting the labels of concepts in both terminologies,
2. The anchoring step first identifying the correspondences between TLAB's entities and the attributes of LOINC concepts and then the correspondences between the TLAB entity and the LOINC concepts.
3. The data-driven validation.

[7] https://fr.wikipedia.org/wiki/Liste_d%27abr%C3%A9viations_en_m%C3%A9decine

Mapping of tokens

In this step, we used the ServoMap tool for the mapping of tokens that constituted the labels of the terminologies (Figure 2.5). The tokenization process consisted in splitting the labels of TLAB and LOINC according to white-spaces and punctuation. We generated a unique code for each token that did not correspond to a stop-word. The codes of LOINC and TLAB tokens were generated using the prefixes "LNCWORD" and "TERMWORD", respectively. The restriction on stop-words was realized using a list of French stop-words [8]. As a result, the cardinality of mappings between TLAB and LOINC tokens has been computed.

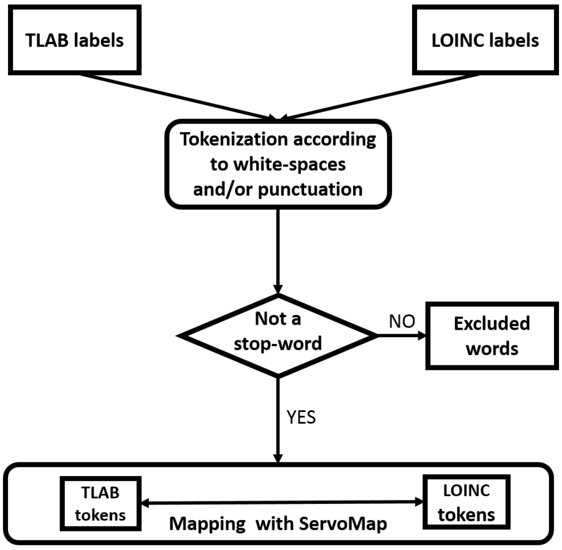


Figure 2.5 – Mapping of TLAB and LOINC tokens starting from their labels.

Anchoring step

The anchoring step consisted in finding similarities between the concepts of the source terminology and the target one. Then, these similarities are validated as equivalence mappings according to the structure of both terminologies. This is a twofold step: (i) the anchoring of TLAB entities to LOINC attributes, and (ii) the anchoring of TLAB entities to LOINC concepts (Figure 2.6).

[8] https://github.com/stopwords-iso/stopwords-fr

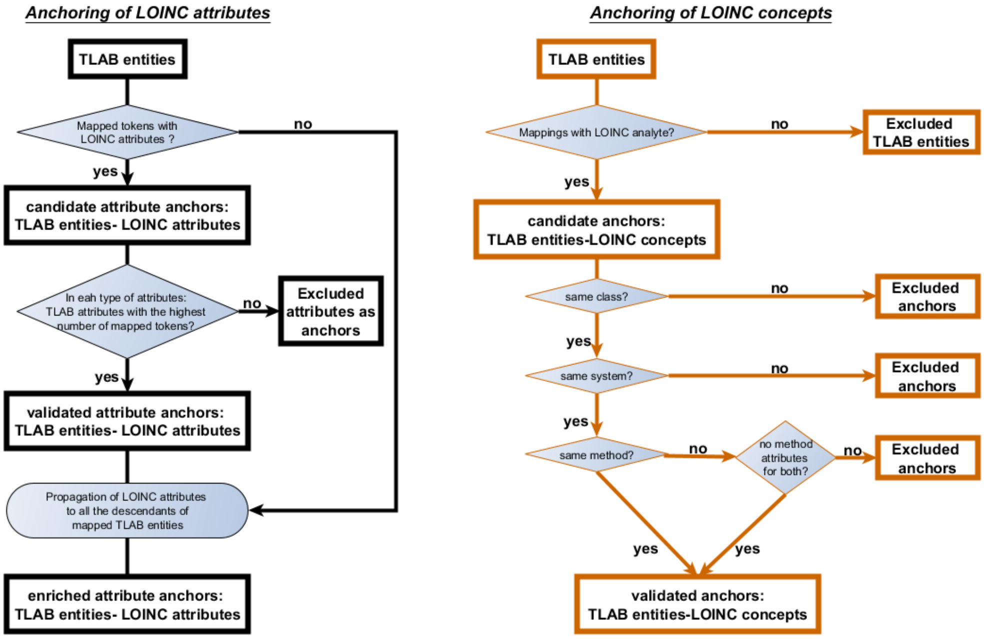

Anchoring of LOINC attributes
TLAB entities
Mapped tokens with LOINC attributes ?
no
yes
candidate attribute anchors: TLAB entities- LOINC attributes
In eah type of attributes: TLAB attributes with the highest number of mapped tokens?
no
Excluded attributes as anchors
yes
validated attribute anchors: TLAB entities- LOINC attributes
Propagation of LOINC attributes to all the descendants of mapped TLAB entities
enriched attribute anchors: TLAB entities- LOINC attributes
Anchoring of LOINC concepts
TLAB entities
Mappings with LOINC analyte?
no
Excluded TLAB entities
yes
candidate anchors: TLAB entities-LOINC concepts
same class?
no
Excluded anchors
yes
same system?
no
Excluded anchors
yes
same method?
no
no method attributes for both?
no
Excluded anchors
yes
yes
validated anchors: TLAB entities-LOINC concepts

Figure 2.6 – Anchoring of TLAB concepts to LOINC attributes and concepts

— *The anchoring to LOINC attributes*

The objective of this stage was to obtain some definitional attributes for TLAB entities. The mapped tokens constituted bridges between TLAB entities and attributes of LOINC concepts. For each type of attributes, when a TLAB entity was mapped to multiple attributes, we chose the attributes having the highest number of tokens in common for the description of this TLAB entity. Then, the attributes related to a TLAB entity were propagated to all its descendants.

— *The anchoring to LOINC concepts*

In this stage, we firstly identified the candidate anchors that correspond to the LOINC concepts and TLAB entities sharing the same analyte. Then, we filtered these correspondences according to class, system and method hierarchies. Thus, the mappings involving entities belonging to distinct classes, systems or methods were deleted. For the last step, when a TLAB entity was not related to a LOINC method, we validated only the anchored LOINC concepts that did not exhibit any method.

Evaluation of the process

To evaluate the process, we proceeded with a data-driven validation process by using lab test results coming from the data warehouse of Bordeaux University Hospital.

The Bordeaux university hospital uses a health data warehouse (based on i2b2[9]) to integrate its data. The data warehouse gathers various structured and unstructured data (clinical data, prescriptions and administration data of medicinal products, biological data, medical imaging data, anatomopathological data and administrative data) for patients who have been visited the hospital at least once since 2010. At the 2019-05-31, the collected data concern 1,591,272 patients corresponding to 11,637,437 visits and 1,152,516,900 observations. Biological data represent 47,9% of all available data (551,823,535 observations).

For the evaluation, we thus used the lab test results that are encoded with TLAB. Test results provide information that is not contained in TLAB labels: the property and the scale of the measurement. The objective was thus to select the anchored LOINC concepts that can be instantiated using the results (*i.e.*, values) associated with each TLAB entity. The evaluation concerned 4,402 TLAB entities that are used to encode 336,758,201 results from (2010-2019) in the laboratory database. From the 4,402 TLAB entities found in the hospital data warehouse, 2,144 could be involved in our process (being non-orphan entities). These entities represented 279,065,808 laboratory results (82.87% of all laboratory results annotated by TLAB entities).

The following three steps were performed for the evaluation process (Figure 2.7):

- We firstly annotated the values and units of measure available in the laboratory database by using the Unified Code for Units of Measure (UCUM[10]) codes. This annotation was realized using a simple morphosyntactic technique for mapping the UCUM code and the units of measure found in the lab results.
- We then manually mapped the properties of the annotated codes in UCUM to the properties of LOINC concepts. Thus, this mapping led to a description of TLAB entities used in the laboratory database with some validated LOINC properties.
- For each TLAB entity, we finally validated the anchored LOINC concepts that could be instantiated using numeric values and also exhibited the same LOINC property. This choice was motivated by the fact that only numeric values were available in the laboratory database. Then, we calculated the

[9] https://www.i2b2.org/

[10] https://unitsofmeasure.org/ucum.html

recall of the validated mappings. Based on these results, we manually curated the 1-1 mappings and computed the precision of results. The validation was realized in a consensual way by two experts with medical and knowledge representation backgrounds. We searched for equivalences between TLAB entities and LOINC concepts or determined if a hierarchical relation existed between them.

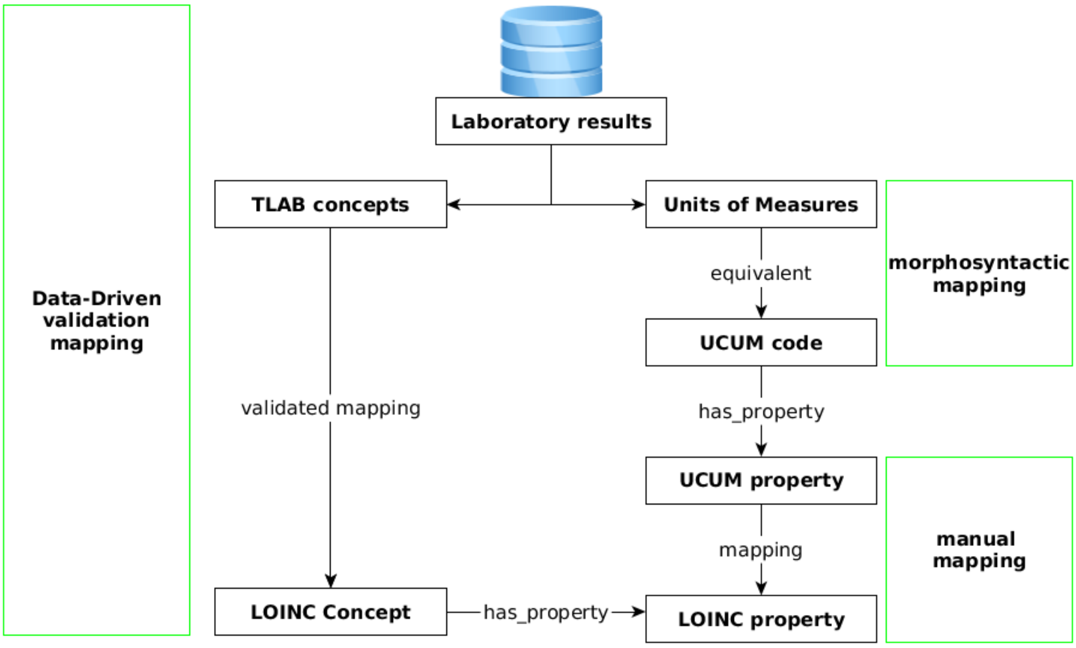


Figure 2.7 – Data-driven evaluation of the mapping strategy between TLAB and LOINC concepts.

2.5 Results

2.5.1 The French structure of LOINC

Two main results presented in this section are the distribution of LOINC concepts according to their related attributes and the distribution of delimited LOINC attributes according to their mappings to the stated attributes.

Firstly, Table 2.1 describes the characteristics of the constructed LOINC structure for the French language, corresponding to 54,480 LOINC concepts, and the stated structure resulting from the LPT content. This table is a quantitative description of the constructed structures of LOINC.

Table 2.1 – Distribution of LOINC concepts according to their relation to LOINC attributes in the French version and the stated structure

LOINC attributes	LOINC French version	Stated structure
Component	22,819	44,313
Analyte	28,807	NA
Challenge	819	1,791
Adjustment	15	35
Property	140	205
Time	31	59
Time aspect	26	NA
Time modifier	3	8
System	394	2,682
Main system	368	NA
Super system	16	62
Scale	6	10
Method	754	1,907
Class	103	389

The stated structure generated much more component, challenge, system, scale, method and class attributes than the French version. For example, the LOINC attribute LP7747-1-- (the dash is actually the label) used as a scale in the stated structure was ignored in our construction process. We found 3,140 components that could be described with a challenge and/or an adjustment. From the computed hierarchy, the process created 28,807 additional analytes. An example of such created analyte from the component label of the LOINC concept 90229-6-*Herpes simplex virus 1 and 2 Ab.IgG and IgM panel - Serum or Plasma* is illustrated in Figure 2.8.

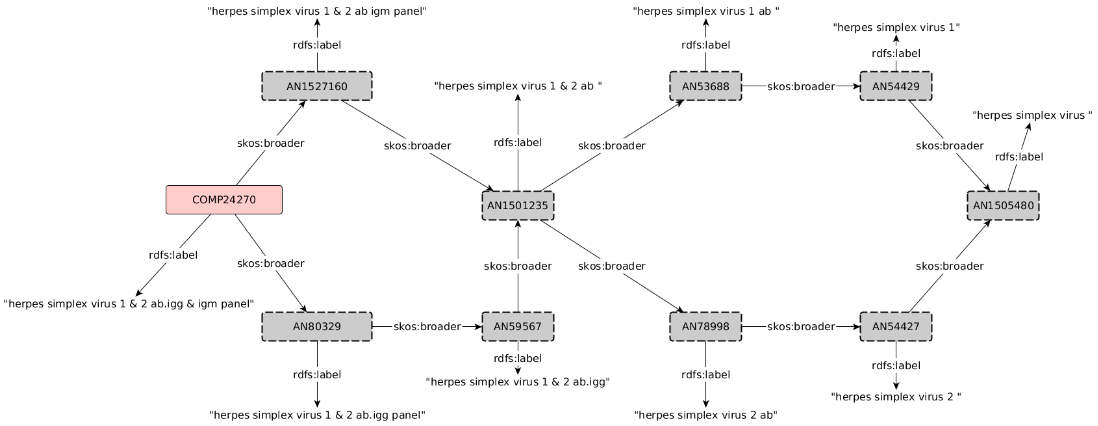


Figure 2.8 – Constructed hierarchy of analytes according to the punctuation in the LOINC component label *herpes simplex virus 1 & 2 ab.igg & igm panel*

Secondly, Table 2.2 describes the cardinality of mappings between the constructed and the stated LOINC structures.

Table 2.2 – Distribution of the constructed LOINC attributes according to the cardinality of their mappings to the stated attributes

LOINC attributes	**1-0 mappings**	**1-1 mappings**	**1-N mappings**
Component	9,710	13,018	91
Challenge	3	804	12
Adjustment	0	15	0
Property	0	136	4
Time	0	31	0
Time modifier	0	3	0
System	43	344	7
Super system	0	15	1
Scale	0	5	1
Method	0	744	10
Class	0	100	3

The mapping process generated a few 1-N mappings. For example, the scale attribute SCALE1503-*quantitatif* was mapped to LP7753-9-*Qn* and to LP7751-3-*Ord* because the LOINC concept 3245-8-*Clot Retraction [Time] in Blood by Coagulation assay* was described with the scale attribute “qualitatif” in the French variant while the other variants used the scale “quantitatif”. This led to erroneously consider that “quantitatif” and “qualitatif” are synonymous labels and to the creation of a unique identifier for them. Another example is the mapping between the component attribute COMP4206-*Sulopenem* and the LOINC attributes LP94456-8-*Linopristin+Flopristin* and LP94455-0-*Sulopenem* because the LOINC concept 55289-3-*Sulopenem [Susceptibility]* has been erroneously described in the Belgian variant using "linopristin+flopristin" as a component while the other variants used "solupenem".

The 1-0 mappings mainly resulted from the disambiguation of multiple attributes mapped to one LOINC concept. The attributes without mappings corresponded to the attributes related through multiples link tags in the stated structure. For example, for the LOINC concept 1352-4-*Yt sup(b) Ag [Presence] on Red Blood Cells from Donor*, the stated structure described it with the following two systems: LP30227-0-*RBC^donor* and LP7536-8-*RBC*. Thus, no mapping could be found because these relations were not computed. On the contrary, our process described the concept with the system attribute SYST3303-*RBC^donor*, itself being related to the super system SSYS1533-*RBC* through a *skos:broader* relation.

2.5.2 Pre-processing of TLAB labels

Figure 2.9 describes the results of the pre-processing of TLAB labels. Overall, they are constituted of 6,593 words. From these words, 1,532 validated NSWs were detected. These NSWs corresponded to 1,105 entry indexes of the dictionary. We manually provided a translation for 191 of them, which impacted 3,005 labels. In addition, the correction using Wikipedia abbreviations involved 1,840 labels. For example, from the label *"R.pcr.S.aureusUri"*, we identified two NSWs for which a correction existed (*i.e.*, **"Uri"**-*"urine/ urinaire"* and **"pcr"**-*"pcr/réaction en chaîne par polymérase"*). The following five labels were thus created:

1. *"R. réaction en chaîne par polymérase S. Aureus Urine"*,
2. *"R. réaction en chaîne par polymérase S. Aureus Urinaire"*,
3. *"R. pcr S. Aureus Urine"*,
4. *"R. pcr S. Aureus Urinaire"* and
5. *"R. réaction en chaîne par polymérase S. Aureus Uri"*.

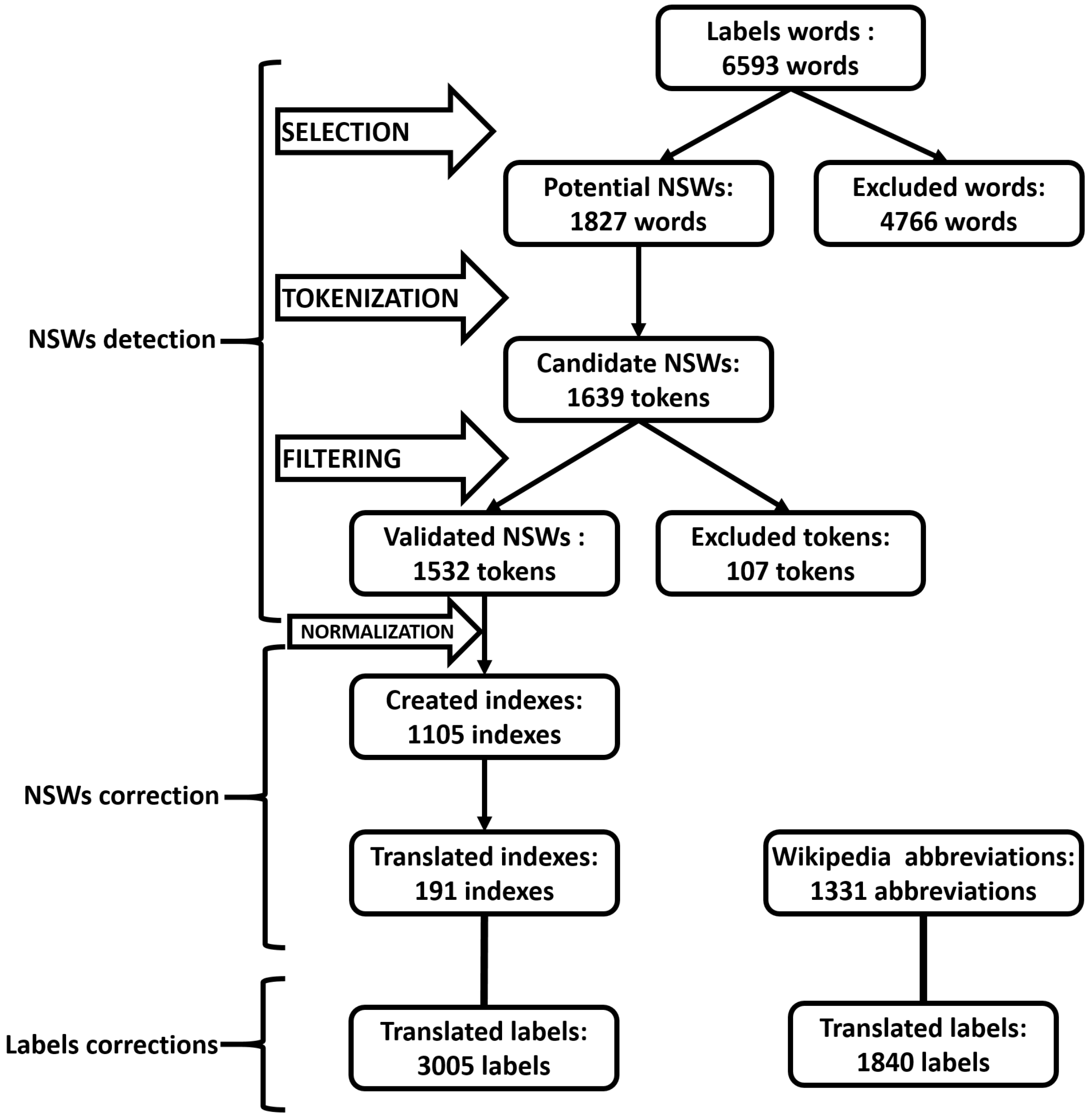

Labels words : 6593 words
SELECTION
Potential NSWs: 1827 words
Excluded words: 4766 words
TOKENIZATION
NSWs detection
Candidate NSWs: 1639 tokens
FILTERING
Validated NSWs : 1532 tokens
Excluded tokens: 107 tokens
NORMALIZATION
Created indexes: 1105 indexes
NSWs correction
Translated indexes: 191 indexes
Wikipedia abbreviations: 1331 abbreviations
Labels corrections
Translated labels: 3005 labels
Translated labels: 1840 labels

Figure 2.9 – Results of the detection of NSWs.

2.5.3 Mapping of tokens

Table 2.3 describes the cardinality of mappings between the tokens of TLAB and LOINC.

Table 2.3 – Distribution of TLAB and LOINC tokens according to the cardinality of resulting mappings

		TLAB tokens	**LOINC tokens**
Mapping cardinality	1-0	2,389	10,327
	1-1	2,226	2,347
	1-N	120	63
Total		4,735	12,737

The tokens that could not be mapped corresponded to 50.45% (2,389/4,735) of TLAB tokens and 81.07% (10,327/12,737) of LOINC tokens. Each LOINC token has been mapped up to two TLAB tokens. Conversely, TLAB tokens have been mapped up to four LOINC tokens. For example, the TLAB token TERMWORD3191-*stimule* has been mapped to the LOINC tokens LNCWORD12088-*stimulating*, LNCWORD3858-*stimul*, LNCWORD7284-*stimule* and LNCWORD10831-*stimulated*. The last LOINC token has also been mapped to the TLAB token TERMWORD2926-*stimul*.

2.5.4 Anchoring step

From the mapping of tokens, we inferred triplets composed of a TLAB entity, an attribute relation and a LOINC attribute. The first inference corresponded to 8,521,727 triplets. These triplets have been reduced to 1,365,129 after considering, for a same type of attribute, the LOINC attributes that share the highest number of tokens with TLAB entities. As an example, for the TLAB entity syn-ana-vtal1-*PCR Adéno/LCR* (with its alternative labels “réaction en chaine par polymérase Adéno / liquide céphalo-rachidien” and “réaction en chaine par polymérase adénopathie liquide céphalo-rachidien”), the algorithm selected SYST1723-*liquide céphalorachidien* rather than SYST1533-*liquide vitrée* because the TLAB entity shared two tokens (*liquide* and *céphalorachidien*) with the first LOINC system attribute, whereas it shared only one token with the second LOINC system (*liquide*).

Anchoring of attributes

Table 2.4 describes the distribution of TLAB entities according to their anchoring to LOINC attributes. By propagating the LOINC attributes associated with each TLAB entity to all their corresponding descendants, almost all TLAB entities have been related to an analyte. The number of TLAB entities that have been related to a LOINC system, a LOINC method or a LOINC class was multiplied by 1.5 (from 3,371 to 5,065 entities), 1.6 (from 4,949 to 7,944 entities) and 3.2 (2,262 to 7,137 entities), respectively.

Table 2.4 – Distribution of TLAB entities according to their anchoring to LOINC attributes

Anchored attributes	Validated anchors	Extended anchors
Component	7,688	8,295
Analyte	7,756	8,295
Challenge	3,362	4,561
Adjustment	348	1,093
Property	1,656	2,374
Time	462	794
Time aspect	456	788
Time modifier	0	0
System	3,371	5,065
Main system	3,213	4,968
Super system	502	767
Scale	139	211
Method	4,949	7,944
Class	2,262	7,137

Anchoring to LOINC concepts

At the beginning of the anchoring process, only five TLAB entities have not been anchored to any LOINC concept and ten TLAB entities have been anchored to a unique LOINC concept. The other TLAB entities have been anchored to multiple LOINC concepts with a maximum of 24,017 LOINC concepts for one TLAB entity. The filtering step based on the LOINC classes, systems and methods reduced the number of mapped LOINC concepts for a TLAB entity. Thus, the filtering step increased the number of TLAB entities anchored to a unique LOINC concept (from 99 (after filtering by class) and 354 (after filtering by system) to 1,011 TLAB entities (after filtering by method)). Concurrently, the increase of TLAB entities anchored to only one LOINC concepts was accompanied by a reduction of TLAB entities anchored to multiple LOINC concepts (from 8,195 (after filtering by class) and 8,149 (after filtering by system) to 6,880 TLAB entities (after filtering by method)).

2.5.5 Data-driven evaluation process

We found that 1,942 TLAB entities were related to 92 units of measures for the description of laboratory results (corresponding to 279,065,424 laboratory results). We mapped 57 units of measures to UCUM codes. These UCUM codes corresponded to 24 UCUM properties, which were mapped to 77 LOINC properties. Thus, 1,187 TLAB entities have been instantiated with 8,455 LOINC concepts. This corresponds to a recall of 0.61.

The median cardinality of mappings was reduced from 20 to 5 LOINC concepts and the maximum from 5,254 to 1,227 LOINC concepts. As an example, for the TLAB entity syn-ana-i261c-c261-*pholcodine*, the LOINC concept 73720-5-*pholcodine ige ab [units/volume] in serum* was selected as the appropriate anchor rather than 81971-4-*Pholcodine IgE Ab RAST class [Presence] in Serum* because the results encoded with syn-ana-i261c are presented with the kUA/L unit of measure. The 1,187 TLAB entities covered 152,159,025 laboratory results. The manual evaluation concerned 197 mappings (1-1 mappings), of which 92 were deemed equivalent and 25 mappings corresponded a subsumption relation. This corresponds to a precision of 0.59.

2.6 Conclusions

In this implementation, we anchored an interface biology terminology to LOINC. The process consisted in taking advantages of the LOINC structure to compensate for the absence of appropriate labels in TLAB for establishing equivalences, or at least hierarchically-related mappings. To deal with the French labels of TLAB, we created a SKOS structure of LOINC integrating the different French variants present in the LOINC release. Thus, this study highlighted the difficulties in involving some interface terminologies in an alignment process because of their noisy labels. In addition, the implementation also showed that the instantiation of terminology entities (*i.e.*, the data encoded with these terminologies) can be used to guide the alignment validation process.

If some authors used machine learning algorithms [142, 161] to deal with noisy labels, we chose a more controllable process by firstly correcting the TLAB labels and the using the semantics of the LOINC structure. Indeed, the performance of machine learning algorithms can be boosted by rich corpora and/or large data source but the labels of TLAB cannot be used to constitute such a rich corpus.

2.6.1 The construction of a formal structure for LOINC

We constructed our own structure of LOINC and we did not use the tables describing the multi-axial hierarchies and the structure of LOINC parts (available in the release) for four reasons. First, the hierarchy table of LOINC concepts is manually maintained and the hierarchy is not meant to describe LOINC as a pure ontology but according to the different domains of laboratory analyses. Secondly, the description of parts is ambiguous. As illustrated in Figure 2.3, multiple LOINC attributes of the same type may be used to describe a LOINC concept. In this example and for a formal description, LP40415-9-*Herpes simplex virus 1+2 Ab pattern* is the appropriate component but LOINC does not prioritize it in the description of the LOINC concept 13505-3-*Herpes simplex virus 1+2 Ab pattern [interpretation] in Serum*. Thirdly, in the multi-axial table, hierarchies are not available for each attribute. In the example of Figure 2.3, no hierarchical link exists between LP40415-9 and the other components. Especially between LP40415-9 and LP14822-8-*Herpes simplex virus 1+2*, a hierarchical relation would clearly be expected (not to say that the "antibodies of herpes virus" are a kind of "herpes virus" but rather to highlight that an analysis on "antibodies of herpes virus" is an analysis on "herpes virus"). Finally, the LOINC attributes are not available in French as elements of the release.

When comparing the obtained structure with the constructed LOINC using the LPT (only for LOINC concepts related to a unique attribute of each type), we noticed some inconsistencies due to translation errors. However, this construction gives a suitable structure for our alignment process. The limitations we highlighted did not question the quality of the constructed structure but gave the possibility to audit the translation process.

2.6.2 The pre-processing of TLAB labels

Enhancing the quality of TLAB was intended to reduce the occurrence of naming conflicts between identical concepts. To correct TLAB labels, we adapted strategies described in the literature for the detection and correction of NSWs used in forums, social networks and SMS [155]. Our strategy was based on punctuation and uppercase, like in [162]. Deliberately, delimited unique letters and words containing integers were not considered as NSWs because many of them corresponded to medical terms (*e.g.*,"B" in "lymphocytes B" and "H1N1" in "virus H1N1"). For this reason, our strategy excluded abbreviations such as *O2T* for "oxygénothérapie" that were consequently not corrected. The preprocessing exhibited reasonable performances with 3,005 labels corrected thanks to corrections applied to NSWs and 1,840 corrected thanks to the Wikipedia dictionary.

This step clearly demonstrates that a standardized process can be applied to improve the quality of noisy labels present in interface terminologies, which are barriers for mapping processes based on morphosyntactic techniques (naming conflicts). In addition, it is a step that is useful locally because it helps to enhance the quality of concept labels.

2.6.3 The alignment process

Our results unsurprisingly highlighted the poverty of overlapping terms between TLAB and LOINC (44% of TLAB tokens and 18% of LOINC tokens have been successfully mapped, as shown in Table 2.3). However, each implemented step in our work tried to overcome semantic conflicts that could occur. To address naming conflicts between identical concepts, our strategy consisted in over-interpreting the mappings between tokens by considering them as sufficient to induce an anchor between a TLAB entity and a LOINC attribute. On the other hand, these over-interpretations led to the occurrence of scaling or confounding conflicts between identical concepts as well as naming conflicts between different concepts. Those negative effects have been partly resolved by the use of the LOINC structure during the filtering step (Figure 2.10).

Combined conflicts between identical concepts have been overcome by propagating the related LOINC attributes of a TLAB entity to all its descendants. To illustrate this last situation, syn-ana-cy301-*soit* has correctly been anchored to 48432-9-*fructose [molar amount] in unspecified time semen* thanks to its hierarchical relation with syn-ana-csfru-*FRUCTOSE SPERME*. Conversely, with the same inaccurate label, the other TLAB entity syn-ana-cy133-*soit* has correctly been anchored to 50193-2-*cholesterol in ldl.narrow density [mass/volume] in serum or plasma* thanks to its hierarchical relation with syn-ana-cldl-*CHOLESTEROL LDL*. Thus, these mappings have been successfully established between entities that did not share the same label or the same definitional elements (these TLAB entities cannot be related to LOINC attributes). These correct anchors illustrated the naming conflicts existing in TLAB.

Finally, we observed that all the characteristics used in the description of LOINC labels cannot be found in an interface terminology label. The main characteristics that can be expected in a TLAB label are the analyte, the system and sometimes the technique. For this reason, only these attributes were used in the mapping process. In addition, the difference of granularity between TLAB concepts and LOINC concepts induced some multiple mappings for some

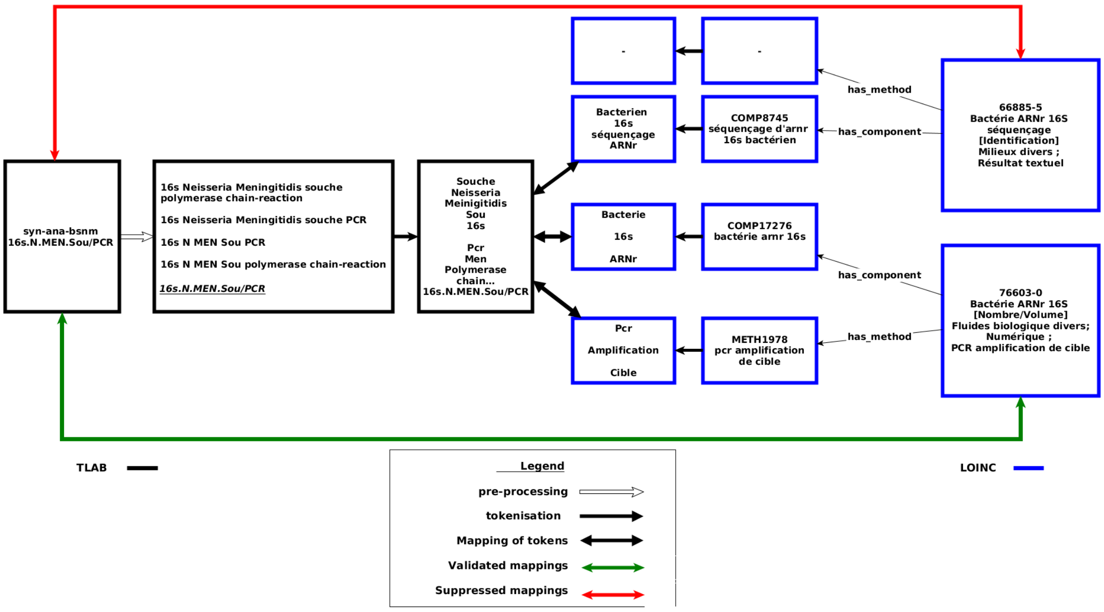


Figure 2.10 – Illustration of the alignment process for the TLAB entity syn-ana-bsnm-*16s.N.MEN.Sou/PCR*: the pre-process step reduces the occurrence of naming conflicts. The filtering step helps to detect and correct cases of confounding conflicts.

TLAB concepts. For example, syn-ana-i202f-*f202 noix cajou* was anchored to 6718-1-*cashew nut ige ab [units/volume] in serum* and 7183-7-*cashew nut igg ab [units/volume] in serum*. Using the original version of LOINC, other authors used the LOINC group structure by seeking the parent concept of the anchored LOINC concepts [142]. However, as illustrated in the previous example, such a parent does not always exist.

We used classical methods for the natural language processing of TLAB and LOINC labels. However, machine learning algorithms could be combined with our process to enhance the quality of results. In addition, our strategy being based on the structure of LOINC, a perspective could be the use of a formal definition and a more formal language, like the Web Ontology Language (OWL [11]) in spite of the SKOS/RDF language used in this work. The appropriate format, integrating all the linguistic variants and all LOINC parts, groups and hierarchical structures (pre-existing or automatically created), could allow to better disambiguate the multiple anchors by choosing the more general one or the parent of all of them.

2.6.4 Focus on semantic conflicts

In conclusion, the proposed strategy is mainly based on the similarity between labels (T) computed thanks to morphosyntactic techniques. However, it also focus on the similarity between interpretants (D) favoured by the support model derived from LOINC. Indeed, through this model, contexts (Co) were provided to entities of both terminologies based on their labels, which enabled to compare the interpretants. Two main aspects regarding the resolution of naming conflicts were thus addressed by our process:

— the **reduction of naming conflicts between identical concepts** by using multiple linguistic variants for LOINC labels and creating alternative labels for TLAB entities,

— the **elimination of naming conflicts between different concepts** by using the structure of LOINC as a support in a repair process.

Thus, the main lesson learned from dealing with naming conflicts is that some conflicts can be directly resolved, but in other cases, the risk of occurrence can just be reduced by an appropriate approach without identifying specifically such conflicts. However, the alignment strategy and performance are oriented according to the identification of equivalent concepts across the knowledge resources without handling links between different concepts. In the next section, difference between concepts is taken into account. We show how to overcome

[11] https://www.w3.org/TR/owl-features/

semantic conflicts by organizing entities describing different notions into a coherent structure.

Let food be thy medicine and
medicine be thy food.

Hippocrates

Chapter 3

Integration process: representation of medicinal products

Summary: In this chapter, we present the work we have done to integrate RxNorm with the part of SNOMED CT that describes medicinal products. The aim was to provide an automatic process for supplying SNOMED CT with RxNorm concepts. In addition, because SNOMED CT has newly adapted its model to comply with international recommendations, this integration was a first step to assess the adherence of RxNorm to international standards for the description of drugs.

Our integration strategy was based on the use of definitional features of entities in SNOMED CT and RxNorm models. We firstly compared these two representation models of medicinal products. We found out that both models shared major definitional features, including ingredient (or substance), strength and dose form. In addition, we highlighted that the representation proposed by SNOMED CT is more rigorous and better aligned with international standards. In contrast, RxNorm describes implicit knowledge, simplifications, and ambiguities in a simpler model.

Secondly, we translated the RxNorm concepts according to the OWL representation of SNOMED CT. Thus, we constructed formal definitions for RxNorm concepts using the ontology design patterns used for describing SNOMED CT concepts. The constructed structure of RxNorm and SNOMED CT have been merged and classified using the ELK reasoner, which highlighted the equivalent concepts between the two knowledge resources as well as the concepts that are specific to each of them.

Finally, the mappings provided by morphosyntactic techniques were compared to the mappings induced by our process (comparison performed according to formal definitions). The divergence between the two approaches showed that our process automatically resolved some naming conflicts. However, our method generated scaling and confounding conflicts. These conflicts were manually identified and corresponded to areas for improvement in RxNorm and SNOMED CT.

Keywords integration process, medicinal products, RxNorm, SNOMED CT, definitional features

Valorization This chapter is the outcome of a project that was realized during an internship at the Cognitive Science Branch of the Lister Hill National Center for Biomedical Communications at the United States National Library of Medicine and supported by the Intramural Research Program of the National Institute of Health. Part of this work has been described in an article entitled "Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability" and published in the proceedings of the 10th International Conference on Biomedical Ontology in 2019. A further paper is in preparation for submission to the Journal of the American Medical Informatics Association.

3.1 Introduction

This chapter addresses the integration process of knowledge resources. In this process, we try to establish equivalences between formal definitions (Df) of entities. We tried to overcome the naming conflicts by using structural techniques. Then, we highlighted manually cases of scaling and confounding conflicts induced by these structural techniques. We applied our integration strategy to the field of medicinal products within which knowledge resources support multiple use cases, such as electronic prescriptions [163–165], drug information exchange, medication reconciliation [166, 167], and data analytics (including pharmacovigilance) [168–170].

The variety of knowledge resources that represent medicinal products induces the need for a formal representation of these entities for facilitating their development and maintenance, as well as for precisely aligning existing drug terminologies [171]. Many definitional characteristics of medicinal products are similar among knowledge resources. For example, clinical drugs are generally defined in terms of ingredient, strength and dose form. However, the level of formalization and the formalism used for representing medicinal products may differ between knowledge resources. Some characteristics may also be specific to some terminologies (especially for country-dependent characteristics, such as the packaging information) [19].

To provide an international framework allowing the interoperability of medicinal product descriptions, international standards have been proposed, such as the Identification of Medicinal Products (IDMP) [20] which is a collection of recommendations from the International Organization for Standardization (ISO). In this work, we tried to integrate RxNorm and SNOMED CT. This choice was largely motivated by the fact that the SNOMED CT recently published a new model for the representation of medicinal products integrating requirements from the IDMP [19]. In addition, SNOMED CT is an international standard being the largest clinical terminology in the world and supported by a consortium of over 40 countries. On the other hand, RxNorm is a standardized nomenclature for the medicinal products used in the United States of America (USA) that has been analyzed [130, 172] and reused to create other standards [173], as well as to integrate drug terminologies worldwide [174]. By integrating these two knowledge resources, a medication list established with RxNorm in the USA could be made available to any electronic health record system in the world, in which drugs are represented using SNOMED CT. In addition, there has not been a detailed comparison between RxNorm and SNOMED CT. Thus, the integration allowing the comparison of both representations could help to improve the structure of these knowledge resources with an international impact.

We firstly describe and compare the models of RxNorm and SNOMED CT [175]. Then, we present the materials that we used and the methods we developed in order to integrate these knowledge resources. Finally, we provide our main results.

3.2 Background

In this section, the models of RxNorm and SNOMED CT are described with a focus on their definitional characteristics.

3.2.1 The SNOMED CT model for medicinal products

The new model of SNOMED CT has been constructed to support international interoperability of medication concepts. For this purpose, the model is restricted to generic drugs and does not represent packs because branded drug names and packages are mainly country-specific [19]. The international goals of SNOMED CT lead it to include the requirements of the ISO which provide the main elements for the description of medicinal products into a set of standards: the IDMP. One principal requirement from IDMP is the representation of clinical drugs in the closed world view. This requirement means that characteristics used to define clinical drugs must be sufficient and that what is not stated is false.

In contrast, according to the open-world assumption underlying OWL, what is not stated is potentially true. For example, the representation of a clinical drug containing "Atorvastatin" must clearly state that it contains "Atorvastatin" as its active ingredient and no other substance being an active ingredient. In the open world view, products containing "Atorvastatin" may also contain other substances, such as "Amlodipine".

Figure 3.1 illustrates the SNOMED CT model describing medicinal products in compliance with the IDMP recommendations. This model is composed of the following six entities, arranged into a subclass hierarchy [1]:

- Two **medicinal product entities**:
    - in the open world view (called medicinal product, or MP): "*A representation of a medicinal product based on description of active ingredients it contains, but not exclusively limited by that description*". An example of a SNOMED CT concept instantiated as a medicinal product in the open world view is 108655000-*Product containing cetirizine (medicinal product)*.
    - in the closed world view (called medicinal product only, or MPO): "A representation of a medicinal product based on description of only and exclusively the active ingredients it contains." An example of a SNOMED CT concept instantiated as a medicinal product in the closed world view is 775140005-*Product containing only cetirizine (medicinal product)*.
- Two **medicinal product form entities**:
    - in the open world view (called medicinal product form, or MPF): "A representation of a medicinal product based on description of active ingredients it contains, but not limited by that description, and on the (generalized) intended site of use for the product." An example of a SNOMED CT concept instantiated as a medicinal product form in the open world view is 768065006-*Product containing cetirizine in oral dose form (me-dicinal product form)*.
    - in the closed world view (called medicinal product form only, or MPFO): "A representation of a medicinal product based on description of only and exclusively the active ingredient(s) it contains and on the (generalized) intended site of use for the product." An example of a SNOMED CT concept instantiated as a medicinal product form in the closed world view is 778701007-*Product containing only cetirizine in oral dose form (medicinal product form)*.

[1] the definitions of each entity are provided in the editorial guideline of SNOMED CT

- One optional **medicinal product containing precisely a given active ingredient** (called medicinal product precisely, or MPP) defined in the closed world view as "a representation of a medicinal product based on description of only and exclusively the precise active ingredients it contains". This optional entity is not currently represented in SNOMED CT but an hypothetical example is *Product containing precisely cetirizine hydrochloride (medicinal product)*.
- One **clinical drug** defined in the closed world view as "a representation of a medicinal product described by its precise active ingredient substances, its manufactured dose form and its strength; strength may be expressed as "presentation strength" or as "concentration strength" as appropriate and the basis of strength substance is explicitly given". An example of such clinical drug is 320818006-*Product containing precisely cetirizine hydrochloride 10 milligram/1 each conventional release oral tablet (clinical drug)*.

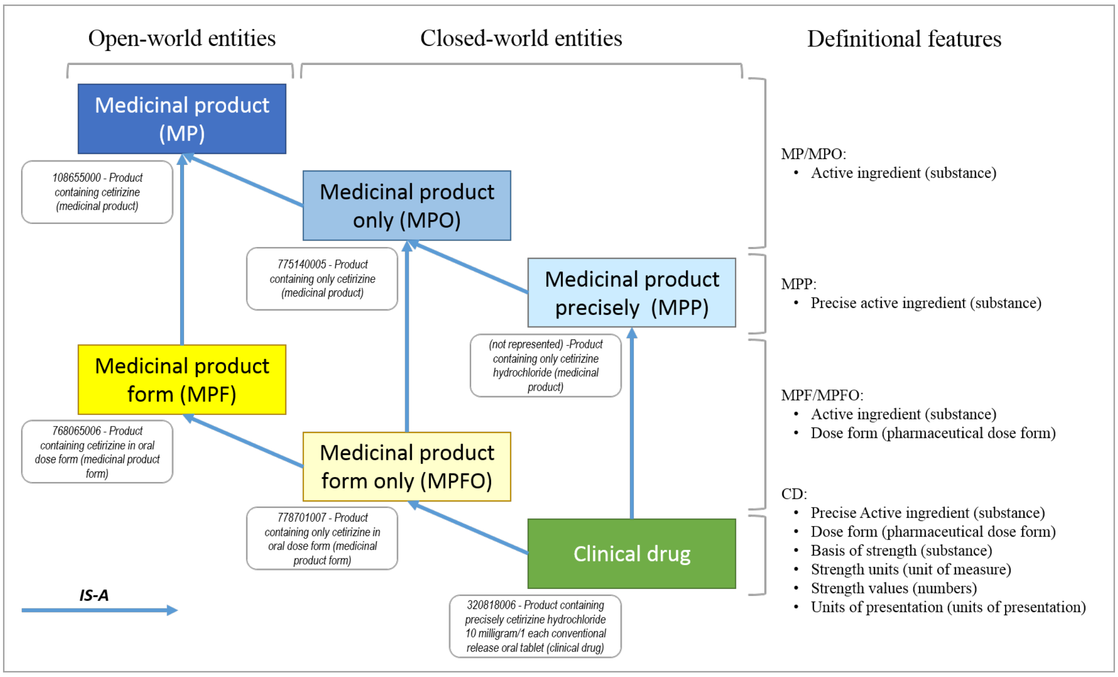

Open-world entities
Closed-world entities
Definitional features
Medicinal product (MP)
108655000 - Product containing cetirizine (medicinal product)
Medicinal product only (MPO)
775140005 - Product containing only cetirizine (medicinal product)
Medicinal product precisely (MPP)
(not represented) -Product containing only cetirizine hydrochloride (medicinal product)
Medicinal product form (MPF)
768065006 - Product containing cetirizine in oral dose form (medicinal product form)
Medicinal product form only (MPFO)
778701007 - Product containing only cetirizine in oral dose form (medicinal product form)
Clinical drug
320818006 - Product containing precisely cetirizine hydrochloride 10 milligram/1 each conventional release oral tablet (clinical drug)
IS-A
MP/MPO:
• Active ingredient (substance)
MPP:
• Precise active ingredient (substance)
MPF/MPFO:
• Active ingredient (substance)
• Dose form (pharmaceutical dose form)
CD:
• Precise Active ingredient (substance)
• Dose form (pharmaceutical dose form)
• Basis of strength (substance)
• Strength units (unit of measure)
• Strength values (numbers)
• Units of presentation (units of presentation)

Figure 3.1 – SNOMED CT model for the representation of medicinal products showing the six types of entities defined in the model along with their definitional features (and their type of values in brackets). Rectangles are model entities and examples of SNOMED CT concepts associated with each type of entity are provided below each rectangle.

The representation of SNOMED CT entities is thus based on "definitional roles" and on related "types of values" as follows:

- **Substance** is the type of values for *active ingredient*, *precise active ingredient* and *basis of strength* (the basis of strength is the substance for which the strength is defined) roles. Examples of substances are 372523007-*Cetirizine (substance)* and 108656004-*Cetirizine hydrochloride (substance)*. All the substances are descendants of the concept 105590001-*Substance (substance)*.
- **Unit of measure** is the type of values for the *strength unit* role, such as 258684004-*Milligram (qualifier value)*.
- **Number** is the type of values for the *strength value* role, such as 3445001-*10 (qualifier value)*. All the numbers are descendants of 260299005-*Number (qualifier value)*.
- **Pharmaceutical dose form** is the type of values for the *manufactured dose form* role, such as 421026006-*Conventional release oral tablet (dose form)*. All the pharmaceutical dose forms are descendants of the SNOMED CT concept 736542009-*Pharmaceutical dose form (dose form)*.
- **Unit of presentation** is the type of values for the *units of presentation* role, such as 732936001-*Tablet (unit of presentation)*. All the units of presentation are descendants of the SNOMED CT concept 732935002-*Unit of presentation (unit of presentation)*.

If there is a top concept for each type of values used as definitional features in SNOMED CT, for the model entities, only the top concept 763158003-*Medicinal product (product)* subsumes all medicinal products entities. However, as highlighted in the examples, semantic tags are used to differentiate the model entities without specification of the appropriate "world view".

In addition, there are no hierarchical relations between substances. However, a *modification_of* relationship may be used to describe the link existing between a modified substance (*e.g.*, ester or salt) and the corresponding base substance (*e.g.*, between Atorvastatin calcium and Atorvastatin). Note that modified substances can be further modified.

Finally, IDMP requires that dose forms be defined in reference to a list of dose forms from the European Directorate for Quality in Medicines (EDQM). EDQM distinguishes between dose forms and units of presentation. Units of presentation are used to express the strength and quantity in countable entities, while dose forms correspond to the physical structure of the medicinal product. In accordance with IDMP requirements, strength units in SNOMED CT are aligned with an international standard for units of measure, the Unified Code for Units of Measure (UCUM). Depending on the unit of presentation, strength can be represented as a concentration strength, a presentation strength or both.

3.2.2 RxNorm model for generic drug

Created in 2002, RxNorm is a normalized terminology for clinical drugs in the USA. RxNorm represents both generic drugs and branded drugs, as well as packs [176]. The full model of RxNorm contains ten entities, five for generic drug entities and five for branded drug entities. For comparison with SNOMED CT, we only present RxNorm generic drug entities and also omit packs.

The simplified RxNorm model for generic drug entities includes the following four entities (Figure 3.2):

- **Ingredient**, including base ingredient (IN), precise ingredient (PIN), and multi-ingredient (MIN) (*e.g.*, IN: 20610-*Cetirizine*, PIN: 203150-*Cetirizine hydrochloride*, MIN: 352367-*Cetirizine / Pseudoephedrine*).
- **Clinical drug component** (SCDC), combining ingredient and strength (*e.g.*, 1011480-*Cetirizine hydrochloride 10 MG*).
- **Clinical drug form** (SCDF), combining ingredient and dose form (*e.g.*, 371364-*Cetirizine Oral Tablet*).
- **Clinical drug** (SCD), combining ingredient, strength and dose form (*e.g.*, 1014678-*Cetirizine hydrochloride 10 MG Oral Tablet*).

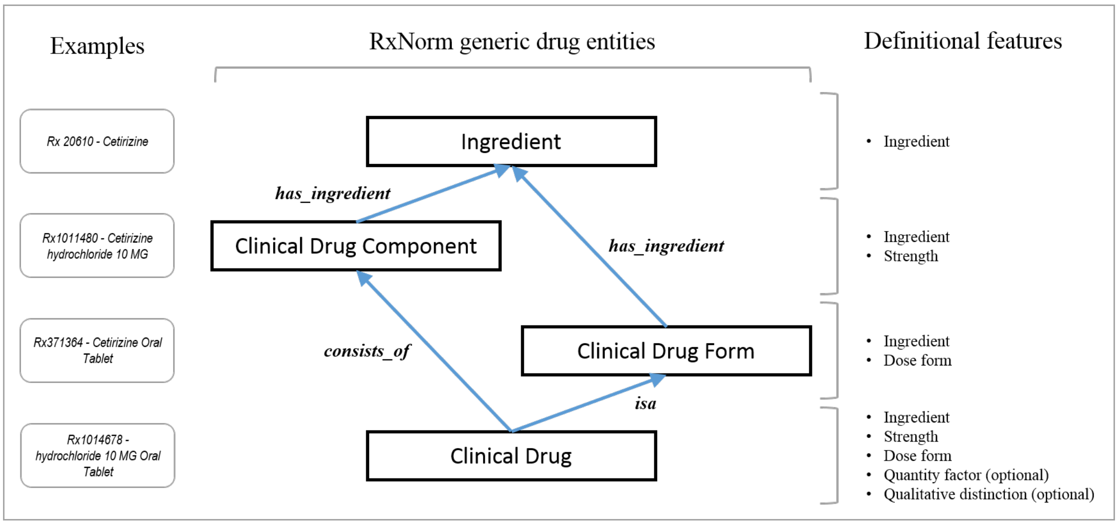


Figure 3.2 – Simplified RxNorm model for the representation of generic medicinal products showing the four types of entities defined in the model, along with their definitional features and examples from the RxNorm knowledge resource.

The representation of these entities relies on three mandatory and two optional definitional features, as follows:

- Mandatory definitional features:
  - ingredient (IN/PIN/MIN). Ingredients in RxNorm can be understood as either the substance contained in a medicinal product, or the class of all medicinal products containing this substance explaining ingredients.
  - dose form (DF) (*e.g.*, 317541-*Oral Tablet*).
  - strength (*e.g.*, *10 MG*).
- Optional definitional features (see below for examples):
  - quantity factor (QF).
  - qualitative distinction (QD).

Strength is normalized in RxNorm. In its units of measure (*e.g.*, for volume, weight, surface), RxNorm uses one unit for each type of quantity (*e.g.*, milligram for weight rather than gram or microgram).

The representation of dose forms in RxNorm is not based on a specific standard [15]. It is also important to note that SCDs and SCDCs refer to the basis of strength substance (*e.g.*, cetirizine hydrochloride), while SCDFs refer to the base ingredient (*e.g.*, cetirizine). Precise ingredients (PINs) generally correspond to modified forms of the corresponding base ingredients (INs). PINs cannot be further modified.

In addition, RxNorm does not explicitly have a notion of "world view" (*i.e.*, open or closed world view) for describing its entities. While clinical drugs implicitly refer to the closed world view, ingredients, clinical drug components and clinical drug forms can be understood in both open and closed world views. Nevertheless, the distinction between the two world views can be assessed through different queries on the RxNorm structure [177].

Finally, the quantity factor (QF) is a number followed by a unit of measure corresponding to vial sizes or patch durations (*e.g.*, "12H"). RxNorm does not explicitly state whether strength is expressed as presentation strength or concentration strength. Presentation strength can be derived from concentration strength by multiplying the concentration strength by the QF (e.g., if the concentration strength is 1MG/ML and the QF is 2ML, the presentation strength is 2MG/2ML). The qualitative distinction (QD) corresponds to qualitative characteristics of a drug different from the main definitional features (e.g., "sugar free" and "abuse-deterrent"). QD and QF are optional modifiers used in RxNorm to define medicinal products when it is clinically relevant to identify such distinctions [15]. All the described entities and definitional features can be accessed through the RxCUI History API [2].

[2] https://rxnav.nlm.nih.gov/RxcuiHistoryAPIs.html #uLink=RxcuiHistory_REST_getRxcuiHistory

3.2.3 SNOMED CT medicinal product design patterns

Ontology design patterns (ODP) are a set of solutions for recurring situations when building knowledge resources. As pointed out in [178], *"ontology design patterns act as an interoperability fallback level through which local conceptualizations can differ to a degree required to appropriately model a given domain or application while still sharing a common conceptual core"*. In other words, when creating locally a design pattern, attention should be paid to the fact that what is designed can be related to a more formal and conventional model.

As a very large knowledge resource, SNOMED CT used ODP to describe its content in description logics (DL) [179]. It contains over 300,000 concepts organized according to a hierarchy rooted by 19 high-level classes. Each SNOMED CT concept has at least one subsumption relation with another SNOMED CT concept.

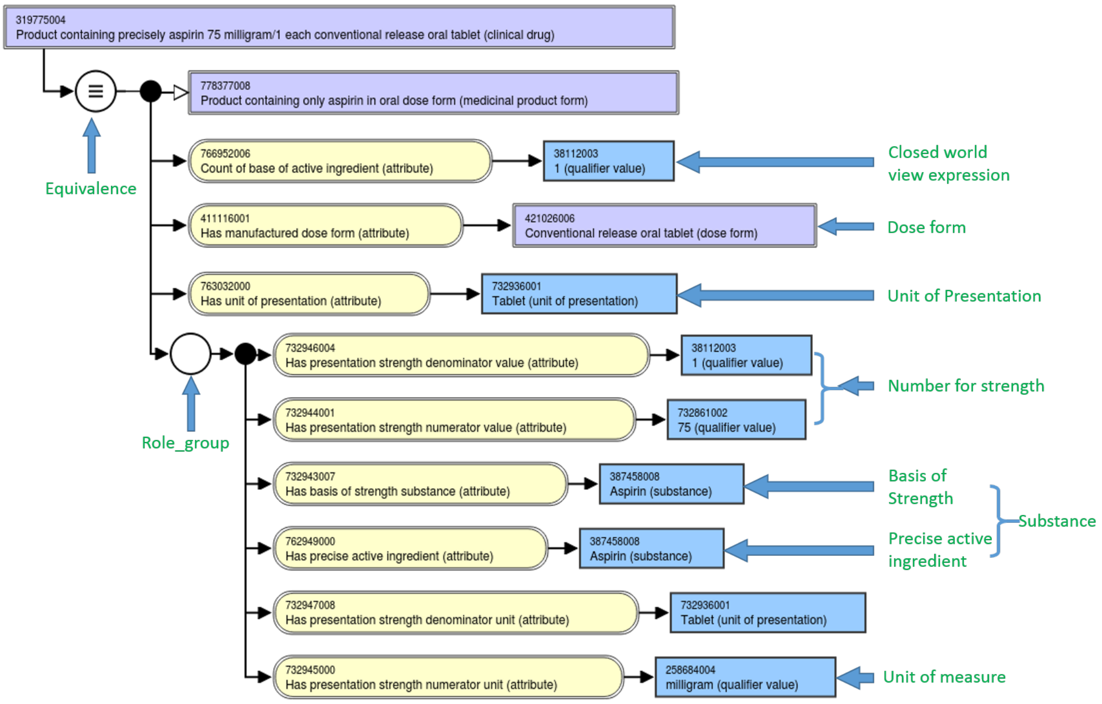

319775004
Product containing precisely aspirin 75 milligram/1 each conventional release oral tablet (clinical drug)
≡
778377008
Product containing only aspirin in oral dose form (medicinal product form)
Equivalence
766952006
Count of base of active ingredient (attribute)
38112003
1 (qualifier value)
Closed world view expression
411116001
Has manufactured dose form (attribute)
421026006
Conventional release oral tablet (dose form)
Dose form
763032000
Has unit of presentation (attribute)
732936001
Tablet (unit of presentation)
Unit of Presentation
Role_group
732946004
Has presentation strength denominator value (attribute)
38112003
1 (qualifier value)
732944001
Has presentation strength numerator value (attribute)
732861002
75 (qualifier value)
Number for strength
732943007
Has basis of strength substance (attribute)
387458008
Aspirin (substance)
Basis of Strength
762949000
Has precise active ingredient (attribute)
387458008
Aspirin (substance)
Precise active ingredient
Substance
732947008
Has presentation strength denominator unit (attribute)
732936001
Tablet (unit of presentation)
732945000
Has presentation strength numerator unit (attribute)
258684004
milligram (qualifier value)
Unit of measure

Figure 3.3 – Description of the SNOMED CT concept 319775004-*Product containing precisely aspirin 75 milligram/1 each conventional release oral tablet (clinical drug)* according to the ODP constructed for the instantiation of SNOMED CT clinical drugs.

Thus, the largest structure of SNOMED CT induces the use of:

— the lightweight *EL*++ [23, 179, 180]: the OWL-EL syntax used for the axiomatic descriptions of SNOMED CT concepts does not contain universal restrictions (*i.e.*, the DL quantifier "∀" (only)). Then, SNOMED CT cannot represent the usual closure axiom to express that a clinical drug is restricted to a given set of active ingredients. Instead, for the description of medicinal products, SNOMED CT adds some axioms of "count of ingredients" to express the closed world view [19].

— the *role_group* relation [181]: it is used to express related description in SNOMED CT (Figure 3.3 [3]). More precisely, this relation is used to describe each characteristic of an active ingredient (*e.g.*, its strength, basis of strength). Thus, for medicinal products containing multiple ingredients, each active ingredient is described with its related characteristics.

Finally, in addition to the "closed world view" needed for clinical drugs, IDMP also required that all the medicinal products be fully described. Thus, unlike the previous version of SNOMED CT (in which medicinal products were mainly primitive concepts), all the medicinal products must be fully defined (*i.e.*, described with equivalence axioms).

3.3 Framework for integrating RxNorm and SNOMED CT

To realize our integration process, it was firstly necessary to compare the two models for identifying a basis for comparison.

We manually searched for equivalences between the entities and between the definitional features of the models. This step was realized on the basis of the definitions of each entity and also based on discussions with the developers of RxNorm and the contributors of the SNOMED International Drug Model Working Group.

First, we disambiguated the notion of ingredient in RxNorm (*i.e.*, IN, PIN, MIN), because it can be understood as either a class of medicinal products (entity) or a substance (definitional feature), as mentioned earlier. Therefore, as shown in Figure 3.4, ingredients in RxNorm correspond to SNOMED CT medicinal products (defined according to open and closed world views) or to SNOMED CT substances, which are active ingredients of SNOMED CT medicinal products.

[3] Annotated diagram from https://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=MAIN&release=&languages=en

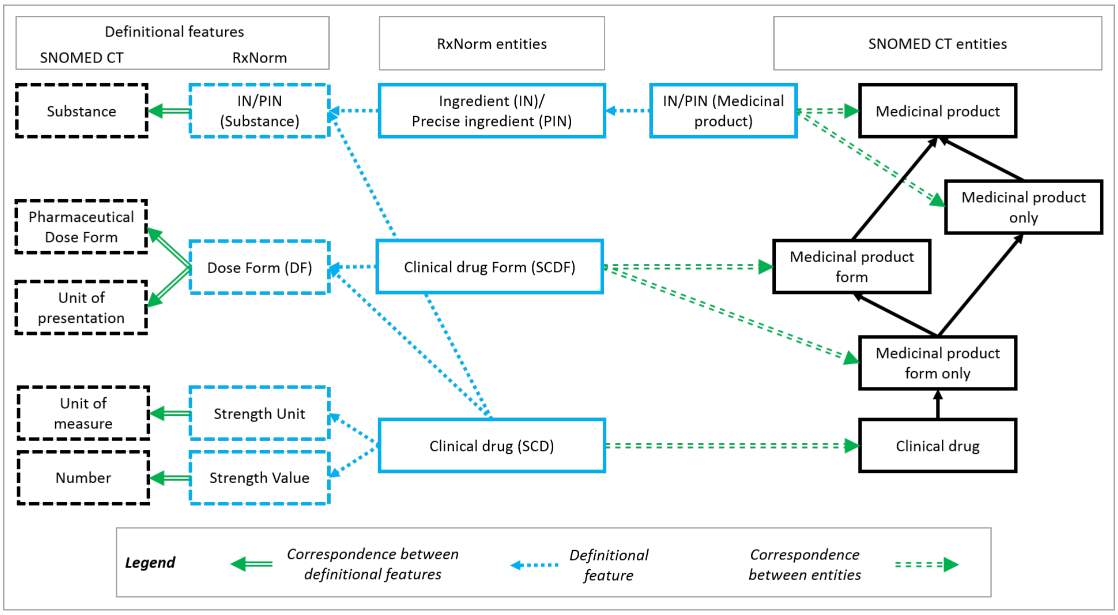


Figure 3.4 – Comparison of RxNorm and SNOMED CT models

RxNorm does not formally have the notion of “unit of presentation”. Units of presentation are implicitly represented through dose forms in RxNorm, whereas the two notions are represented separately in SNOMED CT. For example, in SNOMED CT, “Tablet” is the logical unit of presentation of the conventional release oral tablet, while the two are conflated in the RxNorm dose form “Oral Tablet”. Therefore, RxNorm dose forms generally correspond to pairs of a pharmaceutical dose form and a unit of presentation in SNOMED CT.

In addition, there are no materialized concepts for SCDCs in SNOMED CT. Instead, strength and basis of strength substance are associated as parts of the definition of a clinical drug in SNOMED CT. Therefore, concepts instantiated as SCDCs in RxNorm cannot be related to SNOMED CT concepts, but their defining features are represented as part of clinical drug concepts in SNOMED CT.

SCDs in RxNorm are equivalent to clinical drugs in SNOMED CT as they essentially share the same definitional features. The quantity factor in RxNorm has no direct equivalent in SNOMED CT, but the QF information is implicitly represented in the presentation strength. In contrast, qualitative distinctions are absent from the SNOMED CT model.

While RxNorm only represents one level of modification (between PIN and IN), SNOMED CT can represent arbitrary levels of modification among substances. Both RxNorm and SNOMED CT have the notion of concentration strength and presentation strength. However, RxNorm emphasizes concentration strength (from which presentation strength can be calculated using the quantity factor), whereas SNOMED CT explicitly represents both presentation strength and concentration strength when necessary.

Also, RxNorm normalizes all quantities to one unit (per type of quantity), while SNOMED CT uses units that are most clinically appropriate (following IDMP requirements). For example, RxNorm uses 0.001 milligram (MG) and SNOMED CT uses 1 microgram (UG). This difference merely reflects differences in editorial guidelines, as conversion between the two is trivial.

In conclusion, RxNorm and SNOMED CT models for representing medicinal products are fairly similar and essentially compatible. Both models share major definitional features including ingredients (or substances), strengths and dose forms. Only the qualitative distinction feature of RxNorm has no correspondence at all in SNOMED CT. SNOMED CT is more rigorous and better aligned with international standards. In SNOMED CT, differences tend to be made explicit, *e.g.*, between a substance and the class of medicinal products containing this substance as an ingredient, or between the class of all medicinal products containing only a given substance and the class of all medicinal products containing at least this substance. SNOMED CT also offers more flexibility with relations among substances, as opposed to a fixed precise ingredient to base ingredient relationship in RxNorm. This precision comes at the price of a more complex model, and possibly a steeper learning curve. In contrast, RxNorm contains implicit knowledge, simplifications and ambiguities, but its model is simpler. Based on the comparable characteristics between RxNorm with SNOMED CT models identified in this section, we thus tried to integrate them.

3.4 Materials

To integrate RxNorm with SNOMED CT medicinal products, we used SNOMED CT [4] available in the OWL format (version as of 09/25/2018) and accessed the content of RxNorm through the Restful API [5] (version as of 09/04/2018). Through the Restful API, we also accessed to the mappings provided by RxNorm between its concepts and the SNOMED CT.

3.4.1 SNOMED CT release

The release version of SNOMED CT contains 19,147 classes described as subclasses of 763158003-*Medicinal product (product)*. The structure of SNOMED CT is still under development. In the used version, all concepts were not up-to-date.

[4] https://www.nlm.nih.gov/healthit/snomedct/international.html

[5] https://rxnav.nlm.nih.gov/RxNormAPIREST.html

In addition, entities of the model are not all instantiated. Semantic tags were used for the up-to-date concepts without precision of the world view.

Thus, for the purpose of our work and according to IDMP requirements, we firstly considered only medicinal products that are fully defined. This corresponded to 18,693 SNOMED CT concepts. Based on the ODP described for each entity of the SNOMED CT model, we identified the following number of entities from the model corresponding to SNOMED CT concepts:

— 4,816 SNOMED CT concepts that described medicinal products in the open world view (MP),

— 3,694 SNOMED CT concepts that described medicinal products in the closed world view (MPO),

— 2,725 SNOMED CT concepts that described medicinal product forms in the open world view (MPF),

— 2,609 SNOMED CT concepts that described medicinal product forms in the closed world view (MPFO),

— 4,849 SNOMED CT concepts that described clinical drugs.

The SNOMED CT release did not contain concepts describing medicinal products precisely (MPP).

3.4.2 RxNorm content

The used version of RxNorm contained 18,438 semantic clinical drugs (SCD) to which are associated the following concepts:

— 3,334 ingredients (IN),

— 700 precise ingredients (PIN),

— 1,725 multiple ingredients (MIN),

— 116 dose forms (DF),

— 15,724 clinical drug components (SCDC),

— 8,069 clinical dose forms (SCDF).

All SCD concepts and their linked concepts are uniquely designated by the RxCUI (RxNorm Concept Unique Identifier). The entities of the RxNorm model are considered disjoint (INs are disjoint, DFs are disjoint, etc.)

3.4.3 Asserted mapping between RxNorm and SNOMED CT

SNOMED CT is part of knowledge resources that have been integrated into RxNorm based on a morphosyntactic approach inherited from the UMLS. The mapping between RxNorm and SNOMED CT was extracted from RxNorm (and reflects the US edition of SNOMED CT as of 03/2018). Table 3.1 displays the cardinality and the number of mappings between entities from the RxNorm model and their related SNOMED CT concepts.

Table 3.1 – Mappings between RxNorm and SNOMED CT concepts asserted by RxNorm

		SCD	SCDC	IN	MIN	PIN	SCDF	DF
Cardinality	1-0	12,900	15,723	1,082	1,288	160	8,062	114
	1-1	5,248	1	269	430	456	7	2
	1-N	290	0	1,983	7	84	0	0
Total		18,438	15,724	3,334	1,725	700	8,069	116

These **asserted mappings** involve 14,008 SNOMED CT concepts. These existing mapping were further used to identify relations between the definitional features but also to serve as a gold standard for evaluating the mappings we found.

Unsurprisingly, 99.99% of SCDC have no mapping, confirming that there is no equivalent entity in the SNOMED CT model. The unique mapping was between the SNOMED CT concept 375287000-*Oxygen 100% (product)* and the RxNorm SCDC Rx542303-*Oxygen 100%*, which corresponds to a naming conflict because the SNOMED CT concept represents a clinical drug and not a clinical drug component. Ingredients are mainly involved in 1-N mappings because they are mapped to both medicinal products and substances in RxNorm (*i.e.*, naming conflicts). For example, the RxNorm ingredient Rx83367-*Atorvastatin* has been mapped to the SNOMED CT concepts 108600003-*Atorvastatin (product)* and 373444002-*Atorvastatin (substance)*[6]. Because SNOMED CT does not represent combined substances, multiple ingredients are mainly involved in 1-1 mappings. The low lexical overlap between RxNorm and SNOMED CT for the description of dose forms was highlighted by the almost total absence of mappings for SCDF and DF. Finally, 67% of SCDs were not mapped to any SNOMED CT concept.

3.5 Methods

Three steps have been implemented to perform our process: 1) the disambiguation of RxNorm concepts according to correspondences between the RxNom and SNOMED CT models, 2) the mapping of concepts representing

[6] After the used version for stated mappings, the SNOMED CT concepts have been renamed to better fit the knowledge they represented

definitional features of medicinal products, and 3) the translation of RxNorm concepts.

To evaluate our process, we classified the integrated structure (translated RxNorm concepts and SNOMED CT medicinal products) and compared the inferred equivalences (*i.e.*, equivalences between RxNorm and SNOMED CT concepts that were obtained through classification) with the asserted mappings.

We used the ELK reasoner through the OWLAPI 5.1.0 to analyze the structure of SNOMED CT, to realize the translation of RxNorm concepts and to classify the resulting integrated structure of SNOMED CT and RxNorm. We chose to use ELK because it had been reported that it performs a quick and efficient ranking of SNOMED CT [23, 182].

3.5.1 Disambiguation of RxNorm concepts

This step consisted in mapping the RxNorm concepts to the entities of the SNOMED CT model according to the comparison detailed in subsection 3.3 (*e.g.*, ingredient-medicinal product, semantic clinical dose form-medicinal product form, semantic clinical drug-clinical drug).

For this propose, for each RxNorm concept represented by a RxCUI, we added a semantic tag before each RxCUI corresponding to its related entity in the SNOMED CT model. We thus used “OntoOnlyRx”, “OntoSomeRx”, and “OntoSubstRx”, “OntoDFRx” when the entity describes a medicinal product in the closed world view, a medicinal product in the open world view, a substance and a dose form, respectively.

Strengths in RxNorm were as simple strings. We automatically parsed these strings and generated a list of numbers and units and assigned a unique code to each of them. Each found number and unit was integrated as a subclass of *RxNumber* and *RxUnit*, respectively.

3.5.2 Mapping of definitional features

This steps consisted in establishing mappings between the concepts of RxNorm and SNOMED CT used as definitional features for the description of medicinal products (*e.g.*, ingredient-substance, dose form-dose form/unit of presentation), as follows:

— the mapping of ingredients was obtained from the asserted mappings that are declared in RxNorm. Only mappings involving descendants of the SNOMED CT concept 105590001-*Substance (substance)* were retained.

— the mapping of numbers was realized by a simple morphosyntactic method,

— the mapping of dose forms and units of measure was performed manually.

Table 3.2 – Mapping strategy for RxNorm definitional features and the type of mapping relation

Model entities	**Target hierarchy in SNOMED CT**	**Mapping techniques and strategy**	**Type of mapping relation**
DF	732935002-*Units of presentation*	Manual	Mapping
	736542009-*Pharmaceutical dose form*		Equivalence
Units	767524001-*Units of measure*		
Numbers	260299005-*Number*	Morphosyntactic	
IN/PIN	105590001-*Substance*	Look-up of RxNorm mappings involving substances	

As described in Table 3.2, mappings between RxNorm and SNOMED CT concepts are materialized through equivalence axioms, except for the dose form-unit of presentation mappings that are expressed by a relation called “mappingRelation”. As noticed previously, all RxNorm definitional features being disjoint, multiple mappings were suppressed and only 1-1 mappings were retained (*i.e.*, one SNOMED CT concept for one RxNorm concept, and also one RxNorm concept for one SNOMED CT concept).

3.5.3 Translation of RxNorm medicinal products

For each concept in RxNorm, we defined a generic pattern (*i.e.*, a generic logical definition in OWL), and then we instantiated these patterns for all medicinal products in RxNorm. This step induced an interpretant (*i.e.*, a formal definition (Df) (in OWL)) for each RxNorm concept. Thus, for each RxNorm concept, the pattern was based on the SNOMED CT model for medicinal products and related to its disambiguated concept.

Figure 3.5 illustrates the defined pattern for multiple ingredients as medicinal products in the closed world view.

The description of each pattern is provided in Appendix D. Then, the translated RxNorm concepts and their formal definition are integrated in the OWL format of SNOMED CT. Finally, we classified the obtained structure using the ELK reasoner and selected the concepts of RxNorm and SNOMED CT that were equivalent and denominated them as **inferred mappings**.

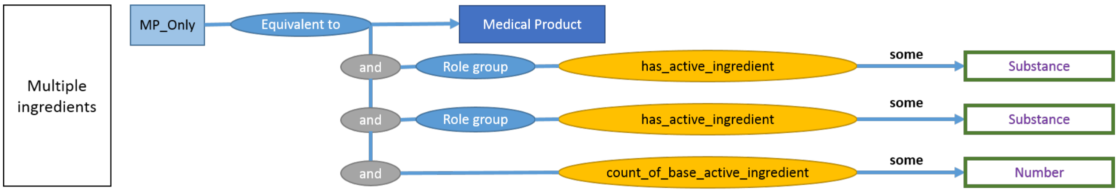


Figure 3.5 – Example of pattern for RxNorm multiple ingredients as medicinal products in the closed world view

3.5.4 Evaluation process

We considered the mappings asserted between concepts in RxNorm and SNOMED CT as the gold standard for the evaluation of the quality of the proposed process. Indeed, the evaluation of this process consisted in comparing the inferred mappings obtained after classification of the logical definitions available for RxNorm and SNOMED CT medicinal products to the mappings asserted in RxNorm. A qualitative analysis of the non-overlapping mappings was realized to analyze the description of medicinal products in both knowledge resources. We performed this evaluation on clinical drugs. Indeed, the SNOMED CT model is currently being instantiated, SCDs have been newly incorporated and were more likely to be conform to the new model than the other entities whose description may be confused with the previous model.

3.6 Results

3.6.1 Disambiguation of RxNorm concepts

RxNorm concepts were disambiguated according to the SNOMED CT model as follows:

- medicinal products in the closed and open world views from INs, PINs or MINs: 5,784 concepts were created for both medicinal product entities,
- medicinal product forms in the closed and open world views from RxNorm SCDFs: 8,286 concepts were created for both medicinal product form entities,
- clinical drug from SCDs: 18,438 concepts were created for clinical drug entities.

As illustrated in Figure 3.6, the RxNorm ingredient Rx42347-*Bupropion* can be used to describe both medicinal products in the closed and open world views (with the appropriate semantic tags used as prefixes).

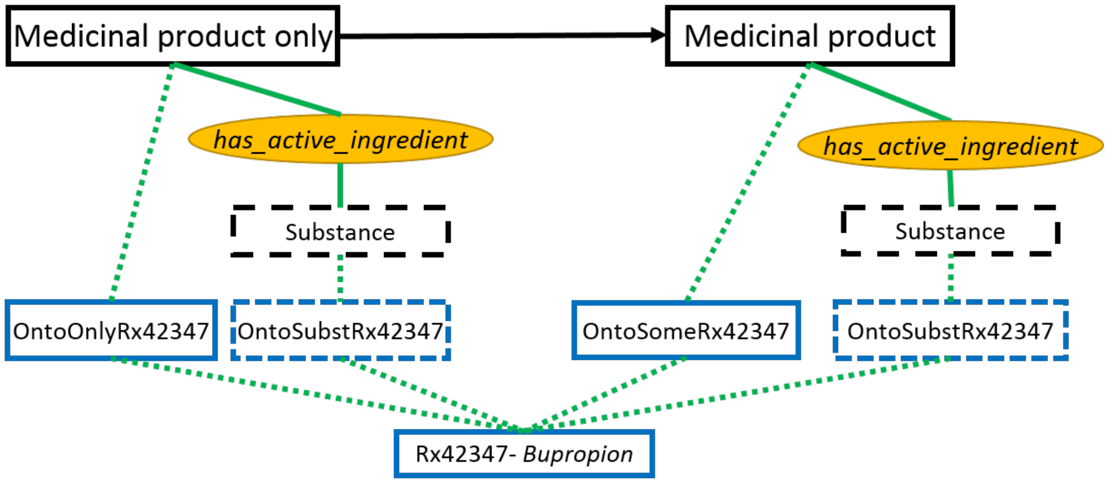


Figure 3.6 – Illustration of the disambiguation of the RxNorm concept Rx42347-*Bupropion* according to the correspondences between the RxNorm and SNOMED CT models.

3.6.2 Mapping of definitional features

Table 3.3 describes the resulted mappings between definitional features in RxNorm and SNOMED CT.

Table 3.3 – Distribution of SNOMED CT concepts according to their mappings to RxNorm definitional features. Source concepts are SNOMED CT concepts, target concepts are RxNorm definitional features and mapped concepts are the SNOMED CT concepts that are involved in the mappings. * 1-N mappings.

Type of mappings	**Source concepts**	**Target concepts**	**Mapped concepts**
Substance-(IN/PIN)	26,743	4,038	3,020
Number-Number	725	1,924	535
Unit of measure-Unit	1,236	18	10
Pharmaceutical dose form-Dose form	307	113	83
Unit of presentation-Dose Form	50	113	*43

The cardinality of resulting mappings is 1-1, except for the mappings between dose forms and units of presentation. An example of such multiple mapping is the SNOMED CT concept 732936001-*Tablet (unit of presentation)*, related to 12 DF in RxNorm, including Rx10312-*Delayed Release Oral Tablet* and Rx970789-*Buccal Tablet*.

3.6.3 Translation of RxNorm medicinal products

All RxNorm ingredients (IN, PIN, MIN) and SCDFs were instantiated using the appropriate patterns. In the process, 1,877 out of 18,438 SCDs were not instantiated because some units of presentation have not been mapped to any dose form. For example, the RxNorm concept Rx763306-*Pantoprazole 40 MG Oral Granules* was not instantiated because no mapping has been found for *Oral granules* with a unit of presentation. Figure 3.7 illustrates the instantiation of the SCD Rx308135-*Amlodipine 10MG Oral Tablet*.

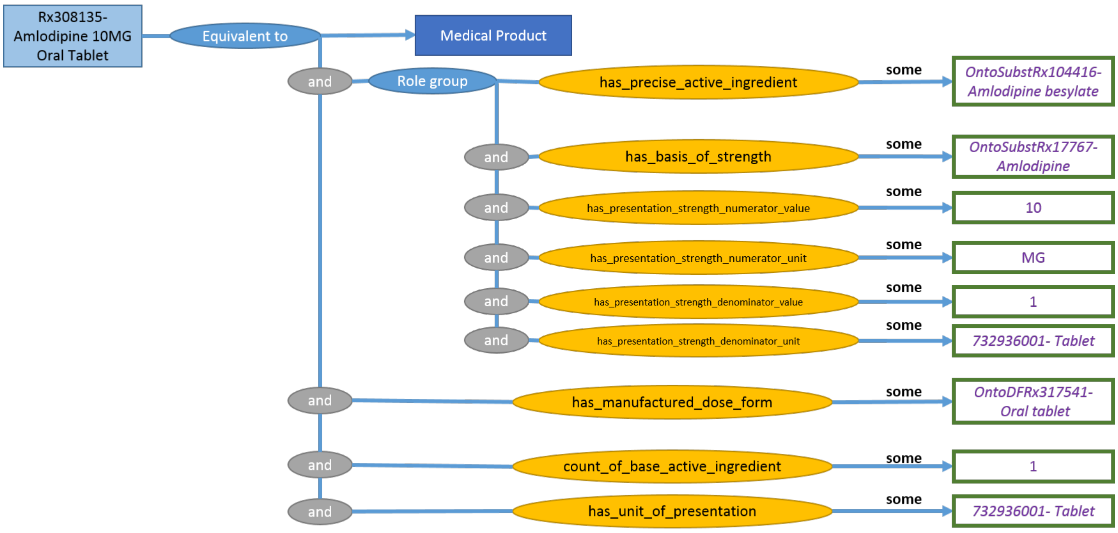


Figure 3.7 – Illustration of an instantiated RxNorm concept (SCD) according to its related pattern

An example of RxNorm concept translation is provided for each type of RxNorm entities in Appendix E.

Table 3.4 describes the inferred mappings between RxNorm and SNOMED CT concepts. For CD, MP and MPO in SNOMED CT, more than 50% of them have been mapped to a RxNorm concept. The results are more contrasted for MPF and MPFO for which only 15 concepts have been mapped to a RxNorm concept.

Table 3.4 – Characteristics of inferred mappings between RxNorm and SNOMED CT concepts. Source concepts are RxNorm concepts, target concepts are SNOMED CT concepts and mapped concepts are the RxNorm concepts involved in the mapping. *Open world view. **Closed world view.

Related entities	**Source concepts**	**Target concepts**	**Mapped concepts**
SCD-CD	16,561	3,204	2,002
Ingredient*-MP	3,299	4,816	2,470
Ingredient**- MPO	3,299	3,694	2,398
SCDF*- MPF	8,069	2,727	15
SCDF**- MPFO	8,069	2,609	15

3.6.4 Evaluation

Focusing on clinical drugs, we found that 11 SNOMED CT concepts were mapped to multiple SCDs. For example, 327082002-*Product containing precisely ciclosporin 25 milligram/1 each conventional release oral capsule (clinical drug)* was inferred as being equivalent to the following two RxNorm concepts: Rx835894-*Cyclosporine, modified 25 MG Oral Capsule* and Rx197553-*Cyclosporine 25 MG Oral Capsule*

Table 3.5 describes the distribution of SNOMED CT concepts according to their inferred and stated mapping with RxNorm. We found that, for 59% (1,892/3,204) of SNOMED CT clinical drugs, the same mappings have been obtained by our process and the morphosyntactic approach used for establishing the asserted mappings. We also highlighted that no mapping could be provided for 9% of SNOMED CT concepts, whatever the applied strategy.

Table 3.5 – Distribution of SNOMED CT clinical drugs according to their mapping to RxNorm concepts: comparison between inferred and stated mappings.

		Asserted mappings		**Total**
		Present	Absent	
Inferred mappings	Present	1,892	110	2,002
	Absent	939	263	1,202
Total		2,831	373	3,204

3.7 Conclusions

3.7.1 Findings

In this chapter, we describe a work that consisted in integrating RxNorm with SNOMED CT using the new SNOMED CT model for medicinal products. We firstly compared the RxNorm and SNOMED CT models and highlighted the definitional features of each entity related to medicinal products. We mapped the definitional features of RxNorm and SNOMED CT and constructed a formal definition (Df) for each RxNorm concept. By classifying the common structure, we found 2,002 equivalences (out of 3,204 possibilities) for clinical drugs between the two knowledge resources (Table 3.5). The applied strategy used both morphosyntactic and structural techniques for identifying similarity between interpretants (D).

The whole process highlighted the compliance of RxNorm with the SNOMED CT model and made RxNorm inherently consistent with SNOMED CT. The process also specifically identified 110 mappings that were not found by the morphosyntactic approach used for establishing the asserted mappings. For these concepts, our structural approach could thus overcome the morphosyntactic limitations (resolution of naming conflicts). However, the structural approach generated some scaling conflicts, with mappings between RxNorm SCDs and multiple SNOMED CT concepts.

The absence of mapping for certain concepts may correspond to the two following situations:

— there exist some errors in the knowledge resources. For example, there is a confounding conflict between 425766008-*Product containing precisely phentermine resin 30 milligram/1 each conventional release oral capsule (clinical drug)* (basis of strength: 426428004-*Phentermine resin (substance)*) and Rx826910-*Phentermine resin 30 MG Oral capsule* (basis of strength: Rx8152-*Phentermine*). According to DailyMed[7], "Phentermine base" seems to be the appropriate basis of strength. Thus, SNOMED CT should use 373343009-*Phentermine (substance)* as the basis of strength. Note that this mapping was established by the morphosyntactic approach but not by our process.

— the granularity of both resources is not always the same. Indeed, a formal definition has been assigned to each of the 16,561 SCDs in RxNorm using the ODP of SNOMED CT. Failing to identify equivalent concepts in SNOMED CT, these concepts and their related definitional features may be

[7] https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=7ca86c66-409b-4852-8631-c3ada6e60738

used to enrich SNOMED CT.

3.7.2 Limitations and perspectives

Firstly, our process did not address some scaling conflicts between the two models. Indeed, 1,877 RxNorm concepts could not be instantiated according to the defined pattern for SCDs because of the absence of units of presentation in RxNorm. For oral solid dose forms, the unit of presentation was used as the denominator unit of the strength. However, when no mapping was stated for RxNorm dose forms, their related SCDs could not be instantiated (see the case of Rx763306 in subsection 3.6.3). In addition, our process did not ensure the conversion of units of measurement, thus inducing scaling conflicts. For example, the equivalence between 326309006-*Product containing precisely desogestrel 150 microgram and ethinylestradiol 20 microgram/1 each conventional release oral tablet (clinical drug)* and Rx249357-*Desogestrel 0.15 MG / Ethinyl Estradiol 0.02 MG Oral Tablet* has not been established by our process. Such missing mapping is consecutive to a granularity difference between the two knowledge resources in the description of strengths.

Secondly, the difficulties for mapping dose forms, which are highlighted by the low coverage of their mappings in asserted mappings (Table 3.1) and in inferred mappings (Table 3.4), need to be overcome. Multiple experts, EDQM[8] as intermediate between RxNorm and SNOMED CT or reverse engineering (using the morphosyntactic mappings between RxNorm and SNOMED CT clinical drugs for inferring mappings of dose forms and units of presentation) are examples of strategies that we plan to investigate in future works.

3.7.3 Focus on semantic conflicts

In conclusion, with a process mainly based on similarities between formal definitions (Df), we highlighted the differences between the morphosyntactic approach (performed in RxNorm) and the structural approach we implemented. We also showed how the comparison of these two approaches can help in the identification of naming, scaling and confounding conflicts. Indeed, the divergence observed between equivalences generated by the morphosyntactic and structural techniques induced three main aspects:

— the **resolution of naming conflicts between identical concepts** by the creation of mappings only available through the equivalence of formal definitions (Df).

[8] https://www.edqm.eu/en/standard-terms-database

— the **elimination of naming conflicts between different concepts** (that are obtained by morphosyntactic techniques) thanks to structural techniques.

— the **occurrence of scaling and confounding conflicts** that can and have to be manually identified.

In the following chapter, we describe how to automatically detect scaling and confounding conflicts and, more importantly, how to automatically overcome them (or, at least, to reduce their occurrence). In addition, we mainly describe the resolution strategy of the last type of semantic conflicts: the **open conflicts**.

Bridges are happy, because they do not judge those who come to them.

Mehmet Murat ildan

Chapter 4

Semantically-enriched integration process: cancer diagnoses

Summary: In this chapter, we were interested in semantically enriching the integration process for being able to link knowledge resources that describe distinct but related domains. We applied the proposed methodology to the oncology field where the reuse of data is confronted with the heterogeneity of knowledge resources. The implemented strategy tried to address all types of semantic conflicts: naming, scaling, confounding and, mostly, open conflicts.

In this frame, we tried to integrate ICD-10 and ICD-O3 by using SNOMED CT as a support. We used two complementary resources (*i.e.*, mapping tables provided by SNOMED CT and the NCI Metathesaurus) in order to find mappings between ICD-10 or ICD-O3 concepts and SNOMED CT concepts. We used the SNOMED CT structure to filter inconsistent mappings (resolution of naming and confounding conflicts), as well as to disambiguate multiple mappings (resolution of scaling conflicts). Based on the remaining mappings, we used semantic relations from SNOMED CT to establish links between ICD-10 and ICD-O3 (resolution of open conflicts).

By creating some complex mappings between ICD-10 and ICD-O3 pairs, we compared the created mappings to the manually performed mappings available in the SEER conversion file and found a recall of 0.50, a precision of 0.68 and an F-measure of 0.58.

The automated process leveraged logical definitions (Df) of SNOMED CT concepts. While the low quality of some of these definitions impacted negatively the semantically-enriched integration process, the identification of such situations made it possible to indirectly audit the structure of SNOMED CT.

Keywords semantic integration, biomedical terminologies, logical definitions, post-coordination, SNOMED CT, ICD-10, ICD-O3

Valorization Chapter 4 is based on the article entitled "Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts" that was published in the Journal of Biomedical Informatics in 2017.

4.1 Introduction

In this chapter, we present our last implementation, which concerned the integration of ICD-10 and ICD-O3, two terminologies maintained by the WHO. Indeed, as previously described (section 1.2.1), these two terminologies need to be integrated in the frame of oncology. The cancer registries, which use ICD-O3, need to incorporate new data that are encoded using ICD-10 from health structures to find the incidence cases of cancer. Also, the registries need to look for cancers' outcome in the health structure database. Thus, the integration of ICD-10 and ICD-03 is fundamental.

However, as described by Jouhet *et al.* [36], ICD-10 and ICD-O3 exhibit structural and semantic heterogeneities. Thus, it is not possible to find equivalences between the concepts of these two terminologies, which correspond to the occurrence of open conflicts. In this context, linking ICD-10 and ICD-O3 requires a true reconciliation of the concepts they describe. Thus, we explored how to automatically resolve each type of semantic conflicts: naming, scaling, confounding and, mainly, open conflicts.

The notion of *reconciliation* in this chapter emphasizes the need for identifying any type of relation that can exist between two concepts (*i.e.*, equivalent and subsumption relations and, in case of disjunction, the appropriate transveral relation), which means to semantically enrich their integration. Through the enrichment of the integration process, we were able to construct some complex mappings [27] between an ICD-10 concept and a post-coordinated expression of two ICD-03 concepts. Complex mappings are correspondences between two elements, in which at least one of these elements is not a single entity but has a more complex structure (*i.e.*, axioms or other expressions).

Thus, our goal was to implement an integration process of ICD-10 and ICD-O3 for a true reconciliation of their concepts, which depends neither on a great structuring of the knowledge resources to be integrated, nor on the expressiveness of their concept labels.

In the following first section, a description of the characteristics of ICD-10 and ICD-O3 is provided before the justification of our methodological choices for performing their semantic integration. Lastly, our analytical framework is featured. In the following sections, we present the materials we used, the methods we developed and the main results of the semantically-enriched integration of ICD-10 and ICD-O3.

4.2 Background

4.2.1 Characteristics of ICD-10 and ICD-O3

ICD-10

ICD-10 is a classification maintained by the WHO for representing nosologic entities through alphanumeric codes. The nosologic entities are autonomous in their determinism. They are also consistent in their clinical manifestations and organized according to their similarities and contrasts. Consequently, ICD-10 concepts are disjoint. Chapter II of ICD-10 is dedicated to tumors, in which 852 alphanumeric codes range from C00 to D48:

- — C00-C97: concepts of malignant neoplasms
    - — C00-C75: concepts of malignant neoplasms, stated or presumed to be primary, of specified sites, except for lymphoid, haematopoietic and related tissues.
    - — C76-C80: concepts of malignant neoplasms of ill-defined, secondary and unspecified sites.
    - — C81-C96: concepts of malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissues.
    - — C97: concept of malignant neoplasms of independent (primary) multiple sites.
- — D00-D09: concepts of in situ neoplasms.
- — D10-D36: concepts of benign neoplasms.
- — D37-D48: concepts of neoplasms having an uncertain or unknown behavior.

The classification of tumors is mainly made by site, and in very large groups, depending on the behavior of the tumor. Each compartment in ICD-10 can be see as a Context (Co) of ICD-10 concepts in the frame of <R,T,D>.

ICD-O3

ICD-O3 is a biaxial classification describing, on the one hand, histological lesions of tumors concepts (morphology), and on the other hand, their anatomical location(s) concepts (topography). The 1032 morphology codes start with the letter "M-" followed five digits between M-8000/0 and M-9989/3. The first four digits represent the specific histologic term , and by extension to the Context (Co) of the morphology ICD-O3 morphology concepts. The fifth digit, behind the slash (/), indicates the behavior of the tumor, *i.e.* whether it is primary malignant (/3), secondary malignant (/6), benign (/0), in situ (/2), with an uncertain or unknown behavior (/9) or undetermined behaviour (/1). The 330 topography codes are composed of four characters and range from C00.0 to C80.9.

4.2.2 Methodological choices for the implemented process

ICD-10 and ICD-O3 are two classifications that differ by:

— The clinical concepts they describe: ICD-10 represents diseases whereas ICD-O3 describes histological lesions and anatomical sites.

— Their structure: ICD-10 is mono-axial while ICD-O3 is biaxial.

This is a typical case of what we called an **open conflict**. Each ICD-10 concept is used independently to record health data and expresses a diagnosis as a whole, thus corresponding to a pre-coordinated concept. In contrast, an ICD-O3 morphology concept must be associated to an ICD-O3 topography concept in order to express the complete diagnosis to be recorded. There are no combination rules in ICD-O3. Thus, all combinations of ICD-03 topography and morphology concepts are potentially allowed. ICD-O3 concepts thus need to be combined for finding mappings with ICD-10 concepts. The link between ICD-10 and ICD-O3 can be made by describing a cancer disease (coded in ICD-10) in terms of its manifestation (ICD-03 morphology concept) and its localization (ICD-O3 topography concept). In a coherent way, the semantic integration of ICD-10 and ICD-O3 firstly required the establishment of the appropriate relation between ICD-10 and ICD-O3 concepts before performing a complex mapping between a ICD-10 pre-coordinated concepts and post-coordinated expressions corresponding to a combination of ICD-O3 morphology and topography concepts.

Because ICD-10 and ICD-O3 are large, their manual reconciliation would be a long and tedious task [183]. On the other hand, morphosyntactic approaches exploiting the concept labels do not take into account the "pre-coordinated" and "post-coordinated" characteristics of ICD-10 and ICD-O3 (cases of naming conflicts). Finally, ICD-10 and ICD-O3 are not described in a formal language. They are not ontologies but just classifications describing disjoint concepts. Thus, integrating them on the basis of their semantic features cannot take into account post-coordination issues. As a result, we used background knowledge available in a support knowledge resource in order to create a method for integrating ICD-O3 and ICD-10. The method using a knowledge resource as a support resource for resource integration is particularly relevant when they are weakly structured or limited to simple classification hierarchies [184]. Indeed, this method allows to compensate this weakness by articulating concepts of knowledge resource to be integrated according to the support knowledge resource, making it possible to automatically find correspondences or to analyze their quality.

4.2.3 Analytical framework

In order to integrate ICD-10 and ICD-O3 by using a support knowledge resource, we have defined a conceptual framework based on the general patterns firstly described by Alekovsky *et al.* [30] and echoed in [184–187] . It consists in two stages (Figure 4.1): **the anchoring stage** (already introduced in section 2.4.3), which aims to generate candidate mappings (called anchors) between concepts of the resources to be integrated and concepts of the support resource, 2) **the derivation stage**, which consists of identifying links between the concepts participating in the anchors within the support knowledge resource so that concepts from the resources to be integrated can be related to each other.

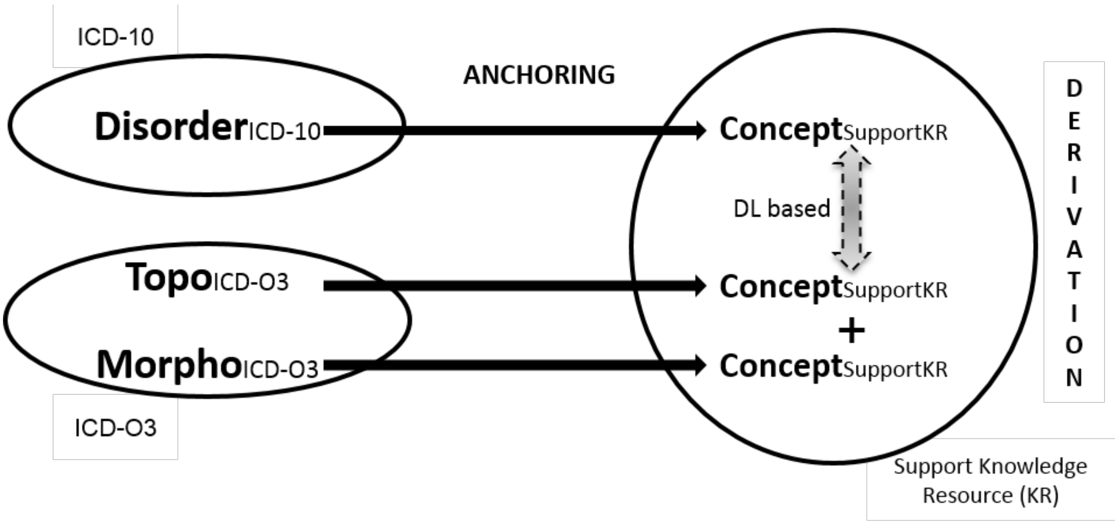


Figure 4.1 – Analytical framework: description of general patterns used to integrate ICD-10 and ICD-O3 using a support knowledge resource. The two stages for this semantically-enriched integration are: 1) the anchoring stage that generates candidate mappings between ICD-10 and ICD-O3 concepts and the support knowledge resource, 2) the derivation stage that identifies links between the concepts of the support knowledge resource which participate in the anchors.

To implement this framework, it was first necessary to select an appropriate support knowledge resource. The latter must be able to describe the appropriate relations between the notions of tumor diseases, histological lesions and anatomical localizations. Its structure must also allow logical inference because ICD-10 and ICD-O3 are large. Indeed, the automatic deduction of relations and constraints existing between the concepts of the support knowledge resource has to be possible, without these relations and constraints being specifically expressed by resources creators [21]. The external resource, which is the most commonly used support knowledge resource within the biomedical domain, is the UMLS Metathesaurus. For instance, it was exploited to align GALEN (Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in medicine) and TAMBIS (Transparent Access to Bioinformatics Information Sources) [188], MedDRA and SNOMED CT [33], as well as CCC (Clinical Care Classification) and NANDA-I (North American Nursing Diagnosis Association-International) [189]. Another biomedical knowledge resource, which has been used as a support for aligning the ATC (Anatomical Therapeutic Chemical) and the MeSH (Medical Subject Headings), is RxNorm [190]. In the specific domain of oncology, the external resource which is often used is the NCI Metathesaurus [191, 192], while Jouhet *et al.* have exploited the NCI thesaurus [193].

4.3 Materials

4.3.1 Support knowledge resource: SNOMED CT

For our study, we chose **SNOMED CT** for integrating ICD-10 and ICD-O3. As one of the most descriptive biomedical knowledge resources, SNOMED CT exhibits ontological characteristics [194] require for our process. SNOMED CT is based on three types of components: (i) *concepts* which represent a clinical meaning and have a unique identifier (SCTID), (ii) *descriptions* which represent labels of these concepts and (iii) *relations* which are binary links between concepts [195].

SNOMED CT also associates logical definitions (Df) to most of its concepts. This logical definition of is composed of other SNOMED CT concepts and relations [196]. In SNOMED CT, it is thus possible to describe a tumor thanks to the semantic link *Associated morphology* relating to a concept describing its histologic lesion as well as the semantic link *Finding site* relating to a concept describing its anatomical location [197]. Through the relation *role_group*, which was introduced in 2002, SNOMED CT could better describe diseases which have several sites or morphological abnormalities. More precisely, the *role_group* relation enables to describe the morphological lesion which is associated with each anatomical site [181, 198]. For example, the logical definition of the SNOMED CT concept 86299006-*Tetralogy of Fallot* (which is a cardiac malformation characterized by different anomalies, which affect multiple anatomical sites) is given in Figure 4.2 [1].

4.3.2 Mapping resources for integrating ICD-10 and ICD-O3

The SNOMED CT mapping tables (SNCTmt). SNOMED CT provides a file which contains, among others, mapping tables between SNOMED CT concepts and ICD-10 as well as ICD-O3 concepts [199]. These mappings have been established manually and their purpose was to find, for a given SNOMED CT code, the corresponding ICD-10 code(s) or ICD-O3 code(s).

The NCI Metathesaurus (NCI Mt). The NCI Mt is a multi-terminology database integrating around 100 biomedical knowledge resources related to cancer [200]. ICD-10, ICD-O3 and SNOMED CT used in this study are included within the NCI Mt. Like in the UMLS Metathesaurus, each concept in the NCI Mt has a unique identifier, named Concept Unique Identifier (CUI), which clusters the codes from distinct knowledge resources supposed to represent the same notion. This clustering has been performed according to a morphosyntactic approach [201].

4.4 Methods

For the semantically-enriched integration process, a preliminary stage has been performed to recover the exhaustive list of ICD-10 codes “from C00 to D48”, as well as the ICD-O3 codes from the NCI Mt. This list has been rid of ICD-10 and ICD-O3 header codes (*e.g.*, C00-C97 Malignant neoplasms) because, in practice, they are not used for diagnostic coding.

As in the previous integration of RxNorm and SNOMED CT, to exploit the structure of SNOMED CT, we applied the ELK reasoner [182]. We used the ELK reasoner trough the OWLAPI 3.5.0 at the anchoring and derivation stages.

4.4.1 Anchoring stage

The anchoring stage consists of three steps: identifying candidate mappings for anchoring, filtering anchors and disambiguating multiple anchors (Figure 4.3). For the two last steps, we used ELK to infer the whole SNOMED CT

[1] Diagram from https://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=MAIN&release=&languages=en

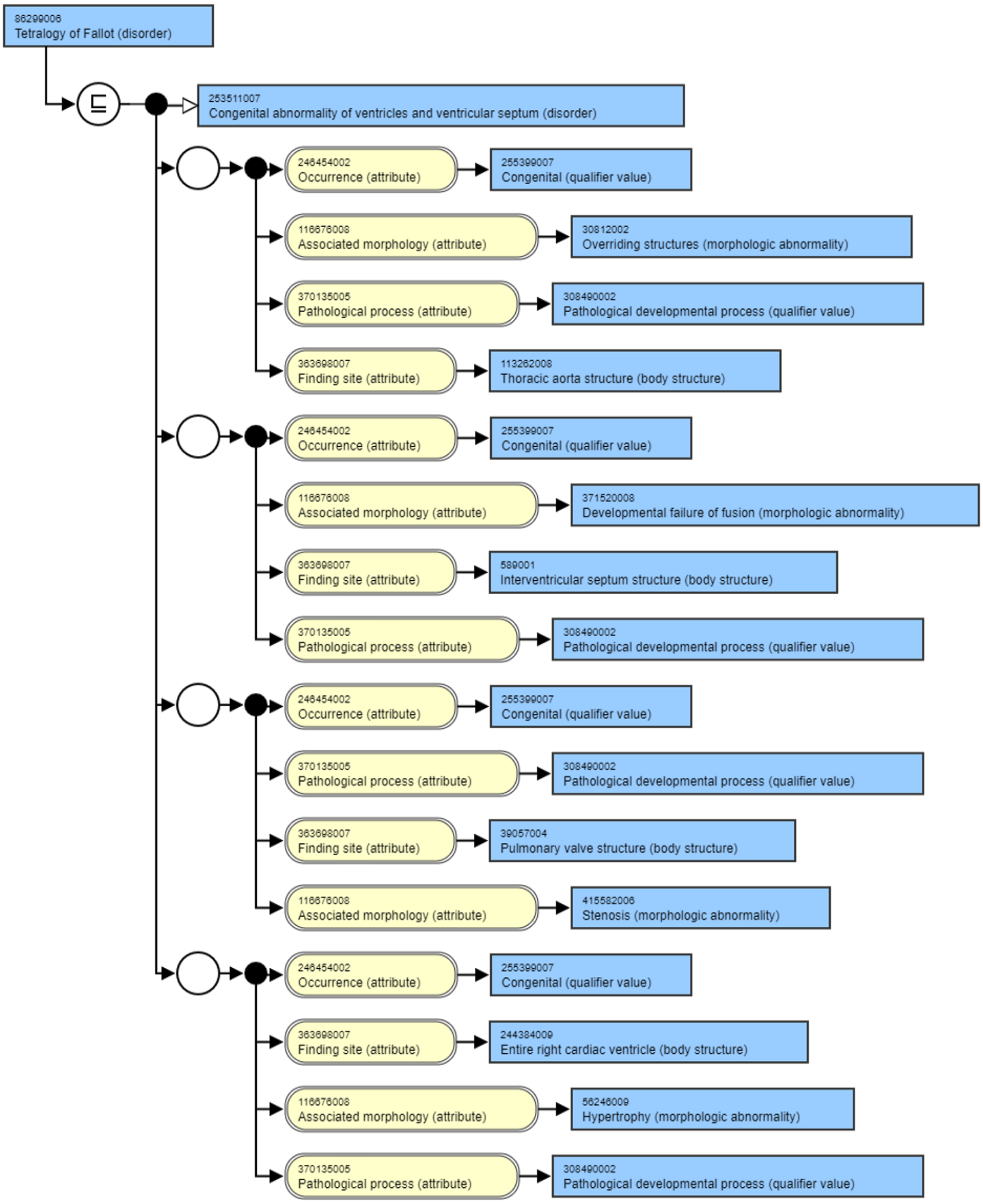

86299006
Tetralogy of Fallot (disorder)
⊑
253511007
Congenital abnormality of ventricles and ventricular septum (disorder)
246454002
Occurrence (attribute)
255399007
Congenital (qualifier value)
116676008
Associated morphology (attribute)
30812002
Overriding structures (morphologic abnormality)
370135005
Pathological process (attribute)
308490002
Pathological developmental process (qualifier value)
363698007
Finding site (attribute)
113262008
Thoracic aorta structure (body structure)
246454002
Occurrence (attribute)
255399007
Congenital (qualifier value)
116676008
Associated morphology (attribute)
371520008
Developmental failure of fusion (morphologic abnormality)
363698007
Finding site (attribute)
589001
Interventricular septum structure (body structure)
370135005
Pathological process (attribute)
308490002
Pathological developmental process (qualifier value)
246454002
Occurrence (attribute)
255399007
Congenital (qualifier value)
370135005
Pathological process (attribute)
308490002
Pathological developmental process (qualifier value)
363698007
Finding site (attribute)
39057004
Pulmonary valve structure (body structure)
116676008
Associated morphology (attribute)
415582006
Stenosis (morphologic abnormality)
246454002
Occurrence (attribute)
255399007
Congenital (qualifier value)
363698007
Finding site (attribute)
244384009
Entire right cardiac ventricle (body structure)
116676008
Associated morphology (attribute)
56246009
Hypertrophy (morphologic abnormality)
370135005
Pathological process (attribute)
308490002
Pathological developmental process (qualifier value)

Figure 4.2 – Tetralogy of Fallot diagram. The circles represent the *role_group* attributes

structure so that subsumption relations which are not explicitly stated between some SNOMED CT concepts are also available.

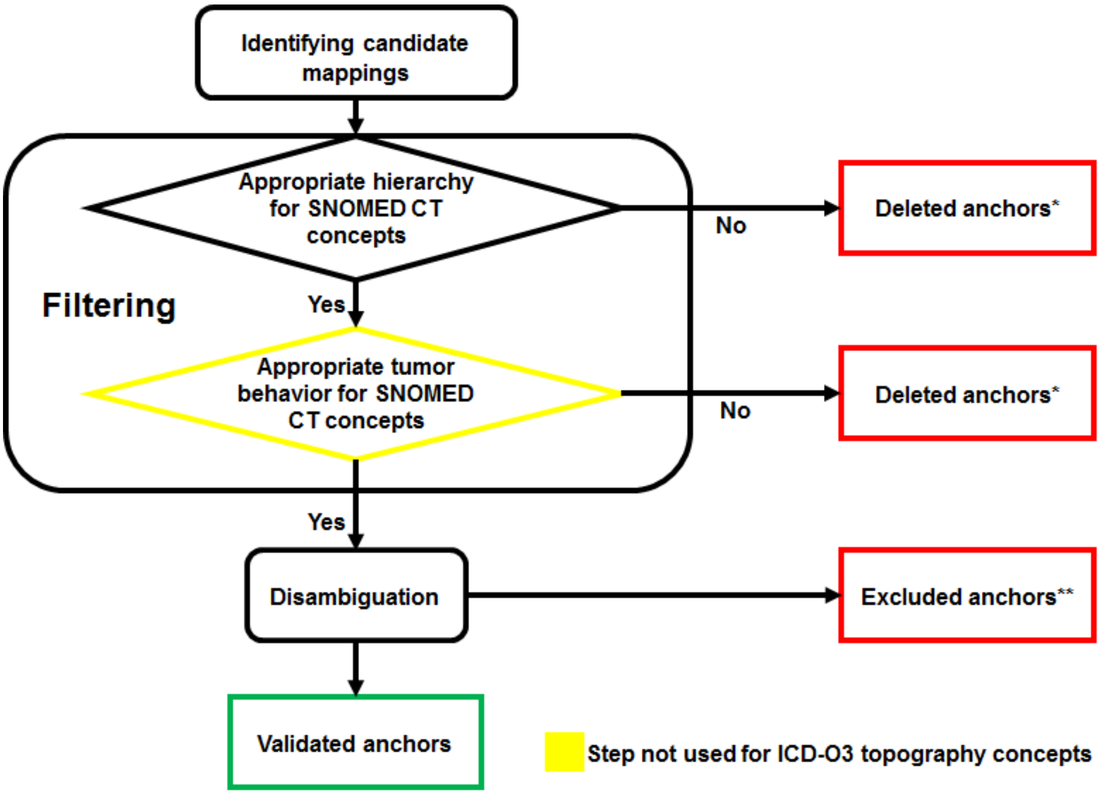

Identifying candidate mappings
Appropriate hierarchy for SNOMED CT concepts
Filtering
Yes
No
Deleted anchors*
Appropriate tumor behavior for SNOMED CT concepts
No
Deleted anchors*
Yes
Disambiguation
Excluded anchors**
Validated anchors
Step not used for ICD-O3 topography concepts

Figure 4.3 – The three steps of the anchoring stage: 1) the identification of candidate mappings, 2) the filtering step, which consists in deleting anchors involving concepts that do not describe the same clinical notions, 3) the disambiguation step for excluding anchors involving a unique ICD-10/ICD-O3 concept and multiple SNOMED CT concepts. *Deleted anchors are erroneous mappings. **Excluded anchors are correct mappings which do not denote equivalences.

Identifying candidate mappings

Two resources were used for the selection of candidate mappings. We first used the SNCTmt by selecting only the anchors described as not obsolete. The second mapping resource, namely the NCI Mt, was exploited to identify the CUI including both a SNOMED CT code (SCTID) and an ICD-10 or ICD-O3 code (Figure 4.4).

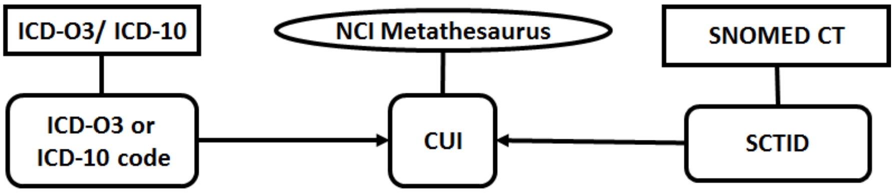


Figure 4.4 – Identifying candidate mappings within the NCI Metathesaurus

Filtering anchors

To eliminate inconsistent anchors, two sub-steps were performed: (i) a filtering according to the SNOMED CT hierarchy and (ii) a filtering according to the tumor behavior. These steps thus consisted in the detection and correction of naming and/or confounding conflicts.

The filtering according to the hierarchy aimed to remove the anchors which involved concepts that do not represent the same general clinical notions. In other terms, it consisted in only validating the mappings between concepts that share the same context (Co). Thus, the anchors were considered as inconsistent in the following cases:

- for ICD-10 concepts (which represent diseases): if the mapped SNOMED CT concept was not a descendant of the concept 64572001-*Disease (disorder)* (The context (Co) of all concepts of disease in SNOMED CT),
- for ICD-O3 morphology concepts (which represent histologic lesions): if the mapped SNOMED CT concept was not a descendant of the concept 416939005-*Proliferative mass (morphologic abnormality)* (The context (Co) of all concepts of histological lesions in SNOMED CT),
- for ICD-O3 topography concepts (which represent anatomical localizations): if the mapped SNOMED CT concept was not a descendant of the concept 91723000-*Anatomical structure (body structure)*(The context (Co) of all anatomic concepts in SNOMED CT).

The filtering according to the tumor behavior was applied only to anchors in which ICD-10 concepts and ICD-O3 morphology concepts participated. This step consisted in the reconciliation of the different classes of tumor behavior found in the structure of the ICD-10 or within the ICD-03 morphology axis with those represented in SNOMED CT. In practice, all anchors involving concepts that do not describe the same kind of tumor behavior were removed. This step was similar to the previous one but it involved more precise (Co) definitions.

Table 4.1 – High-level SNOMED CT concepts corresponding to classes of tumor behaviors in ICD-10 and ICD-O3 : Mapping of the contexts (Co) of ICD-10/ICD-O3 and SNOMED CT

Classes of tumor behavior		**Corresponding SNOMED CT concept(s)**
ICD-10	Primary malignant (C00-C75)	372087000-*Primary malignant neoplasm (disorder)*
	Secondary malignant (C76-C80)	128462008-*Secondary malignant neoplastic disease (disorder)* 302817000-*Malignant tumor of unknown origin or ill-defined site (disorder)*
	Haematological malignancy (C81-C96)	269475001-*Malignant tumor of lymphoid, hemopoietic AND/OR related tissue (disorder)*
	Multiple tumors (C97)	363500001-*Multiple malignancy (disorder)*
	Tumor in situ (D00-D09)	109355002-*Carcinoma in situ (disorder)* 127330008-*Melanoma in situ by body site (disorder)*
	Benign tumor (D10-D36)	20376005-*Benign neoplastic disease (disorder)*
	Unpredictable tumor (D37-D48)	118616009-*Neoplastic disease of uncertain behavior (disorder)*
ICD-O3	Benign (/0)	3898006-*Neoplasm, benign (morphologic abnormality)*
	Undetermined behavior (/1)	86251006-*Neoplasm, uncertain whether benign or malignant (morphologic abnormality)*
	Uncertain or unknown tumor behavior (/9)	6219000-*Neoplasm, malignant, uncertain whether primary or metastatic (morphologic abnormality)*
	In situ morphology (/2)	127569003-*In situ neoplasm (morphologic abnormality)*
	Primary malignant morphology (/3)	86049000-*Malignant neoplasm, primary* (morphologic abnormality)
	Secondary malignant morphology (/6)	14799000-*Neoplasm, metastatic (morphologic abnormality)*

Thus, we have identified the high-level SNOMED CT concepts that correspond to classes of tumor behavior (the appropriate Co for each SNOMED CT concept) which are represented within the ICD-10 structure and within the morphology axis of ICD-O3. The list of high-level SNOMED CT concepts chosen for each class is presented in table 4.1. Some classes of tumor behavior in ICD-10 have multiple corresponding SNOMED CT concepts because these classes represent distinct notions that are not grouped together within SNOMED CT (*e.g.*, the high-level SNOMED CT concepts chosen for the ICD-10 class "Tumor in situ" are 109355002-*Carcinoma in situ (disorder)* and 127330008-*Melanoma in situ by body site (disorder))*.

Disambiguating multiple anchors

The objective of the disambiguation process is to propose the best anchor(s) when several SNOMED CT concepts are mapped to a single ICD-10 or ICD-O3 concept. This means detect and correct the scaling between the mapped concepts. To this end, we examined the existence of subsumption relations between the SNOMED CT concepts. We then kept only the anchor(s) involving the SNOMED CT concept(s) being the most generic (*i.e.*, situated at the highest level in the hierarchy).

Disambiguation does not reduce the number of ICD-10 or ICD-O3 concepts involved in anchors but only the number of SNOMED CT concepts mapped to them. The same disambiguation process was applied to the anchors of each resource (first step), and to the pooling of anchors obtained at the first step (second step).

More precisely, we first addressed the disambiguation of the anchors coming from the SNCTmt independently from those coming from the NCI Mt. This step was intended to harmonize anchors within each of these two resources. At the second step, the disambiguated anchors obtained from the two resources were pooled and a second disambiguation was performed, when needed. Indeed, pooling anchors lead to two situations. Given an ICD-10/ICD-O3 concept:

- Anchor(s) retrieved by the two resources was (were) the same or only one resource retrieved the anchor(s). In this situation, no additional disambiguation was needed.
- Anchors retrieved by the two resources were different (distinct SNOMED-CT concepts). In this situation, the disambiguation process was performed over pooled anchors. Thus, if one of the SNOMED CT concepts involved in the multiple anchors was more general than others, this step allowed to transform a 1-N anchor into a 1-1 anchor.

Evaluation of the anchoring stage

In order to evaluate the methods used during the anchoring stage, we first estimated the coverage of ICD-10 and ICD-O3 concepts within anchors and compared the results obtained through the SNCTmt and the NCI Mt. Then, to assess the impact of each step of the anchoring stage, we calculated the number of anchors obtained for each ICD-10 and ICD-O3 concept and having the following cardinalities before and after each step:

— 1-1 anchors: an ICD-10 or ICD-O3 concept mapped to a single SNOMED CT concept.

— 1-N anchors: an ICD-10 or ICD-O3 concept mapped to more than one SNOMED CT concept.

— 1-0 anchors: an ICD-10 or ICD-O3 concept which could not be mapped to any SNOMED CT concept.

4.4.2 Derivation stage

Derivation method

This step consisted in identifying the relations existing between SNOMED CT concepts participating in the anchors in order to deduce correspondences between ICD-10 concepts and combinations of an ICD-O3 morphology concept and an ICD-O3 topography concept (Figure 4.5). Only 1-1 anchors obtained at the end of the anchoring stage were used for the derivation stage. Therefore, each possible pair of anchored ICD-O3 morphology and topography concepts corresponds to a unique pair of SNOMED CT concepts. For each of these SNOMED CT concept pairs, we looked for the SNOMED CT concepts of disease (equivalent concept or, failing that, parent concepts) that have the appropriate semantic link with each element of the pair (*i.e.*, a *finding_site* relationship with anatomical structures and an *associated_morphology* relationship with histological lesions). The transversal relations, between the related contexts in SNOMED CT for ICD-10 and ICD-O3 concepts, were manually determined. Toward this end, we automatically generated DL-queries which were executed over the inferred SNOMED CT structure obtained with the ELK reasoner. Then, either an equivalent SNOMED CT concept was found or, failing that, the parent concepts of this DL expression were recovered. We finally checked automatically if some of these SNOMED CT disease concepts were anchored to ICD-10 concepts.

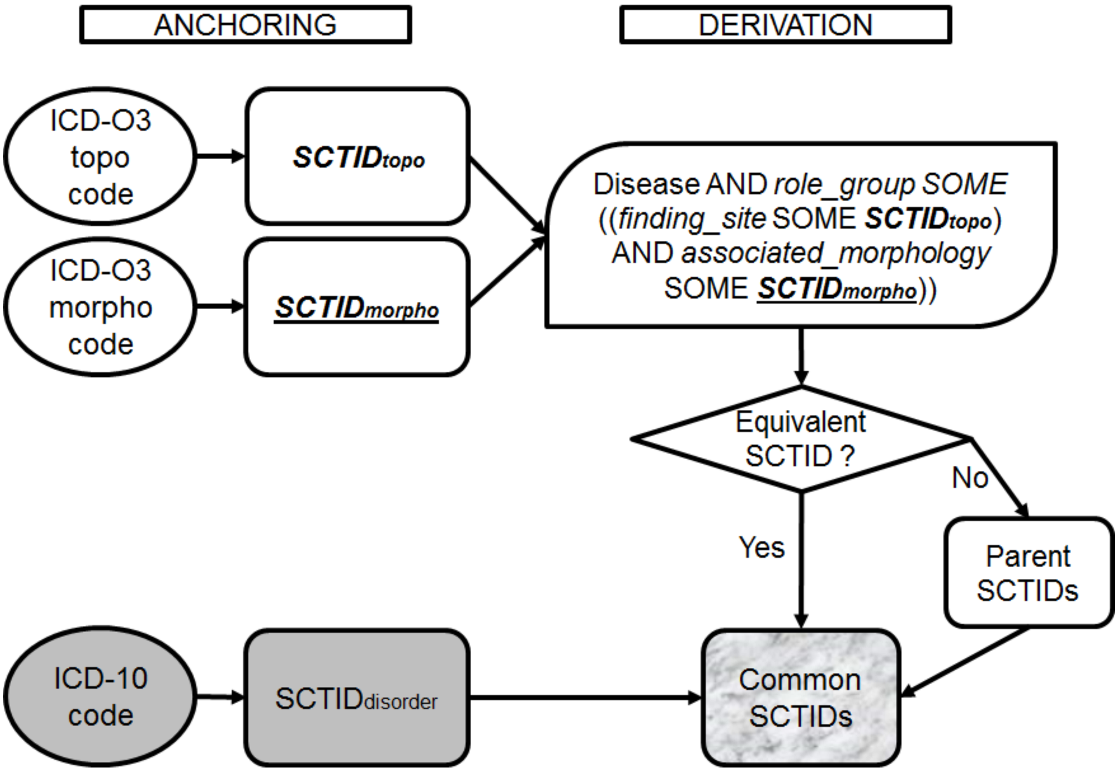


Figure 4.5 – The derivation stage: identifying SNOMED CT concepts of diseases that can be used as a bridge between ICD-10 and ICD-O3 concepts. For each pair of SNOMED CT concepts anchored to ICD-O3 concepts, a DL-query was performed to retrieve the expression corresponding to the disease. The equivalent concept of this DL expression was searched and if it did not exist, parent concepts were used.

Evaluation of the derivation stage

For the evaluation of the derivation stage, we carried out a qualitative and quantitative analyses of the integration results.
For a quantitative analysis, we calculated the number of derivations found for each ICD-10 concept according to the following cardinalities:

- 1-1 derivations: an ICD-10 concept derived with a single pair of ICD-O3 morphology and topography concepts.
- 1-N derivations: an ICD-10 concept derived with more than one pair of ICD-O3 morphology and topography concepts.
- 1-0 derivations: an ICD-10 concept which could not be derived with any pair of ICD-O3 morphology and topography concepts.

We also calculated the coverage of ICD-10 and ICD-O3 concepts involved in the derivation.
For a qualitative analysis, we compared our results with a gold standard, an ICD conversion file provided by the National Cancer Institute within the SEER (Surveillance, Epidemiology, and End Results) program [2]. Within this file, only the correspondences between ICD-10 and ICD-O3 concepts that participated in 1-1 anchors were used for the integration assessment. We thus calculated the overlap of our results with the 23,694 correspondences available in the SEER program conversion file.

4.5 Results

4.5.1 Anchoring stage

Coverage of ICD-10 and ICD-O3 concepts involved in anchors

Figure 4.6 shows the distribution of ICD-10 and ICD-O3 concepts according to the resource used to establish anchors (*i.e.*, the SNCTmt or the NCI Mt). By considering the two resources (*i.e.*, anchors obtained by the SNCTmt, anchors obtained by the NCI Mt, anchors obtained by the SNCTmt and the NCI Mt), we found that more than 88.0% of ICD-10 and ICD-O3 concepts could be mapped to SNOMED CT concepts. For ICD-O3 morphology concepts, the coverage reaches 99.0% (1025/1032). It is noteworthy that for 28.0% of ICD-10 concepts, only one resource provided an anchor to a SNOMED CT concept.

[2] Available at: `http://seer.cancer.gov/tools/conversion/`

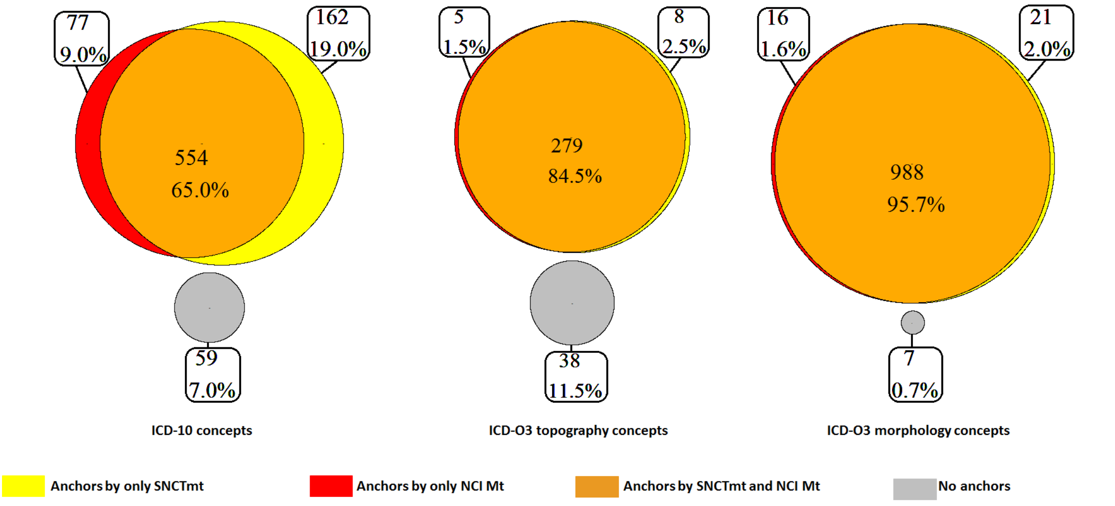


Figure 4.6 – Number of ICD-10 and ICD-O3 concepts involved in anchors, according to the mapping resource used to establish these anchors. The size of circles is proportional to the coverage percentage.

Filtering step

Table 4.2 shows the impact of the filtering process steps, according to each resource used to establish the anchors.

The filtering according to the hierarchy has nearly no impact on the distribution of the ICD-10 and ICD-O3 concepts in the anchors proposed by the SNCTmt. In contrast, in those recovered from the NCI Mt, the number of concepts involved in 1-N anchors decreases; a tendency which is particularly pronounced for ICD-O3 morphology concepts (from 465 to 150) and to a lesser extent for ICD-10 concepts (from 115 to 48). This diminution is accompanied by an increase in the number of ICD-O3 morphology concepts (from 539 to 847) and ICD-10 concepts (from 516 to 572) participating in 1-1 anchors. As an example, within the NCI Mt, the ICD-O3 morphology concept 9684/3-*Malignant lymphoma, immunoblastic, NOS* is anchored to the SNOMED CT concepts 109966003-*Diffuse non-Hodgkin's lymphoma, immunoblastic (disorder)* and 450909005-*Plasmablastic lymphoma (morphologic abnormality)*. The anchor between the ICD-O3 morphology concept (9684/3) and the concept of disease (109966003) was eliminated thanks to the filtering based on the hierarchy. The cardinality of the anchor in which this ICD-O3 concept is involved dropped from 1-N to 1-1. This step thus succeeds in reducing the number of 1-N anchors. On the other hand, some 1-1 and 1-N anchors were eliminated for 11 ICD-10 concepts, 7 ICD-O3 morphology concepts and 29 ICD-03 topography concepts (thus resulting in additional 1-0 anchors).

Table 4.2 – Distribution of ICD-10 and ICD-O3 concepts within anchors obtained by the SNCTmt and the NCI Mt after each filtering step

Steps	**Cardinality of anchor**s	**ICD-10**		**ICD-O3**			
				Topography		Morphology	
		SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Initial	1-0	136	221	43	46	23	28
	1-1	79	516	4	132	960	539
	1-N	637	115	283	152	49	465
Filtering by hierarchy	1-0	136	232	44	75	24	35
	1-1	79	572	4	125	959	847
	1-N	637	48	282	130	49	150
Filtering by tumor behavior	1-0	186	537			72	91
	1-1	159	288			912	838
	1-N	507	27			48	103

The filtering according to the tumor behavior globally leads to a decrease in the number of concepts involved in 1-1 and 1-N anchors, except for ICD-10 concepts with an increasing number of 1-1 anchors coming from the SNCTmt (from 79 to 159). This step results in the elimination of many anchors, in particular for 305 ICD-10 concepts participating in anchors obtained within the NCI Mt. As an illustration, the anchor between the ICD-10 concept C47.3-*Malignant neoplasm of peripheral nerves of thorax* and the SNOMED CT concept 188325002-*Malignant neoplasm of peripheral nerve of thorax (disorder)* was deleted. According to the SNOMED CT hierarchy, this concept is described as being a tumor which can be primary or not, contrary to the ICD-10 concept which is exclusively primary. Although both concepts have the same label, they do not describe the same tumor behavior and, thus, cannot be mapped to each other (case of naming conflict).

Disambiguation step

The number of disambiguated concepts (*i.e.*, whose cardinality of anchors was initially 1-N and became 1-1), respectively mapped through the SNCTmt and the NCI Mt, are 289 and 14 for ICD-10 concepts, 127 and 59 for ICD-O3 topography concepts, and finally 43 and 41 for ICD-O3 morphology concepts (Table 4.3). An example of disambiguation is the ICD-10 concept C50.1-*Malignant neoplasm of the central portion of the breast*, which was initially mapped to the three following SNOMED CT concepts: 93745008-Primary malignant neoplasm of central portion of female breast (disorder), 708921005-*Carcinoma of central portion of breast (disorder)* and 448436006-*Sarcoma of central portion of female* *breast (disorder)*. The disambiguation process was able to detect that among these three concepts, the concept 93745008 being the most general, it was a valid mapping for C50.1. That typically corresponded to the detection and correction of scaling conflicts induced by the manual mapping process.

Table 4.3 – Disambiguation of anchors coming from the SNCTmt and the NCI Mt. 1-0 anchors do not appear because their number is not changed by the disambiguation step.

		Cardinality of anchors	**ICD-10**		**ICD-O3**			
					Topography		Morphology	
			SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Steps	Before*	1-1	159	288	4	125	912	838
		1-N	507	27	282	130	48	103
	After*	1-1	448	302	131	184	957	879
		1-N	218	13	155	71	3	62
Total			666	315	292	255	960	941

Pooled anchors of the SNCTmt and the NCI Mt resulted in the increase of ICD-10 and ICD-O3 participation in anchors. More precisely, the remaining anchors involved 706 ICD-10 concepts, 969 ICD-O3 morphology concepts and 289 ICD-O3 topography concepts. At the end of the disambiguation process, 57.2% (487/852) of ICD-10 concepts, 38.5% (127/330) of ICD-O3 topography concepts and 87.3% (901/1032) of ICD-O3 morphology concepts participated in 1-1 anchors.

4.5.2 Derivation stage

Quantitative analysis

Table 4.4 presents the number of ICD-10 concepts which could be derived with one or multiple pairs of ICD-O3 topography and morphology concepts and those which could not be derived at all. ICD-10 concepts were mainly derived to multiple pairs of ICD-O3 topography and morphology concepts (22.5% for 1-N derivations against 1.3% for 1-1 derivations). An example of 1-1 derivation is D13.2-*Benign neoplasm of duodenum* (ICD-10 concept) with 8850/0-*Lipoma, NOS* (ICD-O3 morphology concept) combined to C17.0-*Duodenum* (ICD-O3 topography concept). This kind of derivation overcame open conflicts and corresponded to a complex mapping. Of note, there were more 1-1 derivations between ICD-10 concepts and pairs of ICD-O3 concepts for the category "Haematological malignancy", probably because haematological tumors are very specific lesions.

Table 4.4 – Distribution of ICD-10 concepts derived with 0, 1 or many pairs of ICD-O3 topography and morphology concepts. *N is the number of ICD-10 concepts of each category

ICD-10 concepts	N*	**Cardinality of ICD-10 concepts derived with pairs of ICD-O3 topography and morphology concepts**						**Total**	
		1-0		**1-1**		**1-N**			
Benign tumor	180	57	31.6%	0	0.0%	34	18.9%	91	50.5%
Haematological malignancy	92	24	26.1%	10	10.9%	20	21.7%	54	58.7%
Unpredictable tumor	86	29	33.7%	0	0.0%	20	23.2%	49	57.0%
Tumor in situ	66	26	39.4%	1	1.5%	14	21.2%	41	62.1%
Primary malignant	388	133	34.3%	0	0.0%	99	25.5%	232	59.8%
Secondary malignant	39	14	35.9%	0	0.0%	5	12.8%	19	48.7%
Multiple tumors	1	1	100.0%	0	0.0%	0	0.0%	1	100.0%
Total	852	284	33.3%	11	1.3%	192	22.5%	487	57.1%

Overall, by combining 1-1 and 1-N derivations, we found that 23.8% (203/852) of ICD-10 concepts could be derived with 38.5% (127/330) of ICD-O3 topography concepts and 86.0% (892/1032) of ICD-O3 morphology concepts.

Qualitative analysis

We found 63,142 ICD-O3 pairs which could be derived with ICD-10 concepts after the derivation process. Among them, 57,505 pairs were each derived with one ICD-10 concept and 5,637 pairs were each derived with multiple ICD-10 concepts. A total of 17,474 ICD-O3 pairs were common with the 23,694 pairs described in the SEER conversion file and for 11,932 of them, our integration process found the same ICD-10 concept as the SEER conversion file. This corresponds to a recall of 0.5; a precision of 0.68 and an F-measure of 0.58. As an example, C15.9-*Esophagus, NOS* and 8504/2-*Noninfiltrating intracystic carcinoma* were derived with D00.1-*Carcinoma in situ of esophagus* both in the SEER conversion file and according to our derivation process.

For the remaining ICD-O3 pairs, our derivation process found different ICD-10 concepts than the SEER conversion file proposes. The ICD-O3 pair formed by C00.0-*External upper lip* and 8856/0-*Intramuscular lipoma* illustrates such cases. Our process resulted in derivations with D10.0-*Benign neoplasm of lip* and D17.0-*Benign lipomatous neoplasm of skin and subcutaneous tissue of head, face and neck* while the SEER conversion file describes a derivation with D17.9-*Benign lipomatous neoplasm, unspecified*.

4.6 Conclusions

4.6.1 Findings

Our study consisted in integrating two biomedical terminologies that focus on diagnostic coding in the field of oncology. The difference of clinical notions represented in ICD-10 and ICD-O3 could not result in 1-1 mappings between their concepts because they are disjoint (*i.e.,* open conflicts). Thus, we did not perform an alignment of these two terminologies but their integration by linking concepts through non-hierarchical relations. Thus, we implemented a semantically enriched integration process, by proposing a method for establishing appropriate transversal relations between ICD-10 and ICD-O3 concepts that resolves cases of open conflicts. We finally identified complex mappings between ICD-10 pre-coordinated concepts (diseases) and ICD-O3 post-coordinated expressions (combinations of topography and morphology). Even if they describe disjoint concepts, these terminologies are organized according to a coherent main classes that was used in our study for their integration as a context (Co) for each of their concept. We chose SNOMED CT as a support knowledge resource for this semantic integration not only because its domain coverage includes those of ICD-10 and ICD-O3 but also because it benefits from ontological characteristics which allowed logical inferences over its structure. Logical definitions (Df) in are based on OWL-EL, which is a *"trimmed down version of OWL that trades some expressive power for the efficiency of reasoning"*. This is the reason why we chose ELK for reasoning over the structure of SNOMED CT, which has previously been shown to be sufficient to express this knowledge resource [202]. In our work, reasoning and DL-queries enabled to retrieve links that were not explicitly stated within the SNOMED-CT structure. Moreover, although built expressions based on ICD-O3 combinations could refer to anonymous classes (because not explicitly described within SNOMED CT), we were able to classify them and link them to an ICD-10 code.

Anchoring stage

By using the SNCTmt and the NCI Mt, we were able to obtain a high coverage of ICD-10 and ICD-O3 concepts within anchors. Thanks to the combined use of the two resources, we indeed found anchors for more than 88% of ICD-10 and ICD-O3 concepts. The highest coverage (99%) concerned ICD-O3 morphology codes, which can be explained by the fact that ICD-O3 morphology concepts were used as support for the representation of SNOMED CT histological lesions [203]. It is noteworthy that, although the overlap is important between anchors obtained by the SNCTmt and the NCI Mt, it was useful to make use of both of these resources because some anchors were found in only one of them, especially for ICD-10 concepts (19% for the SNCTmt and 9% for the NCI Mt).

The main benefit from the anchoring stage was not to create anchors but rather to improve their quality by detecting and resolving semantic conflicts induced by the original mapping strategy. Although ICD-10 and ICD-O3 are poorly structured, we successfully made corrections and reconciled proposed anchors by using their structure at the filtering and disambiguation steps. These steps can thus be qualified as alignment repair processes [37]. The filtering step indeed enabled to delete anchors involving concepts that do not describe the same clinical notion, mainly corresponding to naming conflicts in NCI Mt and confounding conflicts in SNCTmt. The disambiguation step managed to exclude anchors when a hierarchical relationship existed between SNOMED CT concepts involved in multiple anchors, due to scaling conflicts induced by manual mapping in most cases, so that only the most relevant anchor was retained. Thus, these processes highlighted and succeeded in solving the limitations of the morphosyntactic method used by the NCI Mt for establishing mappings and those of the manual method used for creating the SNCTmt. It is important to note that these two methods are the most commonly used in the literature to create mappings, like in systems described previously such as AROMA [204], ServOMap [18] and Onagui [205]. Thus, our methodology may be applied to improve the quality of mappings created by any such application. Indeed, our method is independent of strategies used for creating mappings, because it is only based on the structure of SNOMED CT, ICD-10 and ICD-O3.

Derivation stage

- **Derivation strategy:** In the derivation process, we looked for equivalent, and parent if necessary, concepts of the DL expression corresponding to a pair of ICD-O3 concepts. ICD-10 represents nosologic entities and an ICD-10 concept can represent one or more entities. The notions represented by a combination of ICD-O3 concepts may correspond exactly to the nosologic entity represented by the ICD-10 concept, in which case an equivalence can be found. In contrast, the ICD-O3 combination may represent a nosologic entity which is part of a group of entities represented by an ICD-10 concept. In this situation, subsumption relations are thus of interest.
- **Derivation coverage:** We were able to derive 86% of the ICD-O3 morphology concepts, 36% of the ICD-O3 topography concepts and 24% of the ICD-10 concepts. The coverage of ICD-10 concepts is correlated with the coverage of ICD-03 topography concepts because ICD-10 concepts related to cancer diagnoses are grouped according to the anatomical localization of the tumor. Thus, the absence of anchors for a given ICD-O3 topography

automatically implies the absence of anchors for the ICD-10 concepts involving this anatomical localization. Conversely, the coverage of ICD-O3 morphology concepts is high. This can be explained by the facts that: i) the same histological lesion may exist for different anatomical localizations, and ii) the description of histological lesions in ICD-O3 is more precise than in ICD-10. This difference in the level of precision also explains the numerous 1-N derivations.

— **Derivation quality:** The derivation stage enabled to find an ICD-10 concept for 74% (17,474/23,694) of ICD-O3 pairs of the SEER conversion file. Moreover, our integration process correctly and automatically generated 50% of the correspondences between an ICD-10 concept and a pair of ICD-O3 concepts described in the SEER conversion file.

A potential explanation of the divergences observed between our derivation process and correspondences proposed by the SEER is that its conversion file is based on rules of cancer registries. Conversely, our derivation process intends to relate ICD-O3 combinations to ICD-10 concepts based on their semantics. As a result, our process can find multiple derivations for a single combination whereas the SEER proposes only one of them. For instance, in the SEER conversion file and according to our derivation process, the ICD-O3 pair C75.3-*Pineal gland* and 9769/1-*Immunoglobulin deposition disease* was integrated with D47.9-*Neoplasm of uncertain or unknown behaviour of lymphoid, haematopoietic and related tissue, unspecified* (according to the rule 4.1 of cancer registries for recording an haematopoietic disease [206]). However, our derivation process also proposed D44.5-*Neoplasm of uncertain behavior of pineal gland* for this pair. Although the later derivation is significant, it has not been retained by the SEER. This finding highlights that our process does not depend on specific conversion rules, but only on the semantics provided by SNOMED-CT.

Another consequence of our process was the derivation of pairs that are not medically relevant. An example of such irrelevant combinations is the ICD-O3 pair C50.2-*Upper-inner quadrant of breast* and 8153/1-*Gastrinoma, NOS* which was integrated with the ICD-10 concept D48.6-*Neoplasm of uncertain or unknown behaviour of Breast*. Indeed, "gastrinoma" is a specific morphologic abnormality of the digestive tract so this tumor cannot appear with breast as a primary site. Confronting derivation results with data from cancer registries is a perspective that would allow keeping only the ICD-O3 pairs that are effectively used in practice to record health data. However, it is necessary to underline that our derivation process takes the imperfect but informative coding that may exist in real data (*i.e.*, coding error).

4.6.2 Integration process and evaluation limitations

Our integration of ICD-10 and ICD-O3 concepts remains incomplete. The main limitation of our methods concerns 1-N and 1-0 anchors, which were not derived. For 1-N anchors, the disambiguation process needs to be improved. Some 1-N concepts were still present after the disambiguation of results obtained by the SNCTmt and the NCI Mt but others were also created when pooling anchors coming from these two resources. In many cases, this is a consequence of concepts that are incorrectly represented as siblings within the SNOMED CT structure. As an example, the ICD-O3 morphology concept 8831/0-*Histiocytoma, NOS* was involved in a 1-1 anchor with 128741006-*Deep histiocytoma (morphologic abnormality)* according to the SNCTmt and in a 1-1 anchor with 302843004-*Histiocytoma (morphologic abnormality)* according to the NCI Mt. By pooling the anchors of the two resources, 8831/0-*Histiocytoma, NOS* finally participated in a 1-N anchor because the two concepts 128741006-*Deep histiocytoma (morphologic abnormality)* and 302843004-*Histiocytoma (morphologic abnormality)* are erroneously described as siblings in SNOMED CT (typical case of scaling conflict). Indeed, they must clearly be related through a subsumption relationship. Other hierarchical and transversal semantic links must be sought by the disambiguation process because SNOMED CT apparently does not contain appropriate links between some of its concepts. Therefore, a potential strategy for improving the disambiguation process would be to search for such semantic links in other knowledge resources. As an example, the Foundational Model of Anatomy (FMA) [207] may be a good candidate to identify relations between SNOMED CT concepts which are anchored to a given ICD-O3 topography concept.

4.6.3 Comparison with previous works

The most similar previous work compared to our study is the one realized by Jouhet *et al.* [193], who also tried to integrate ICD-10 and ICD-O3 thanks to a support knowledge resource, namely the NCI thesaurus. If we compare their results with ours, we derived 888 ICD-O3 morphology concepts against 860 for them. By contrast, as our model only considers 1-1 anchors for the derivation stage, the high proportion of 1-N anchors for ICD-O3 topography concepts leads to a lower coverage of ICD-O3 topography concepts (127) and ICD-10 concepts (203) involved in derivations compared to the coverage obtained by Jouhet *et al.*, being respectively 278 and 302. Thus, one of our prospects is the fusion of our results with those obtained by Jouhet *et al.* We would like to check if the use of the NCI thesaurus could improve our semantic integration. In particular, for concepts having no anchors with any SNOMED CT concept, such concepts could be mapped to NCI thesaurus concepts. Finally, we believe that merging the results of both studies will highlight complementarities of the NCI thesaurus and the SNOMED CT. We chose SNOMED CT because it has been used to support the semantic integration of various biomedical knowledge resources in previous works. For instance, Brown *et al.* [208] used it to align the Veterans Benefits Administration (VBA) disability code set and the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9CM). In this work, authors used a morphosyntactic method and compared two approaches. The first one was a direct alignment of VBA and ICD-9CM without using any other resource, while the second approach used the SNOMED CT as a support knowledge resource. The use of SNOMED CT increased the alignment coverage, which illustrates that SNOMED CT is able to cover various clinical domains of medicine. This work differs from ours in that authors did not use the structure of the SNOMED CT to find semantic links between disjoint concepts or to ensure the quality of the mapping process. Finally, the repair process performed in this study was made manually by domain experts in a consensual way. Another example is the work of Bakhshi-Raiez *et al.* [209], who used SNOMED CT to align APACHE II and APACHE IV, which are two versions of a classification system used to encode the reasons for intensive care admission. Firstly, authors manually created mappings between SNOMED CT concepts and those of APACHE II and IV. Then, authors used the SNOMED CT structure to retrieve SNOMED CT concepts which had hierarchical links (especially the *part_of* relationship) with already mapped concepts. Thus, the common SNOMED CT concepts mapped to APACHE II and IV concepts constituted the bridges between the two classifications. As in our study, authors used the structure of SNOMED CT to establish anchors between two knowledge resources but they did not have to realize a semantic integration of disjoint concepts. A challenge raised by the integration of ICD-10 and ICD-O3 was to align pre-coordinated concepts with post-coordinated expressions. To address this issue, Dhombres *et al.* [119] have implemented a strategy requiring that one of the two knowledge resources to be aligned must have sophisticated labels and the other one must be able to carry out post-coordination. Because ICD-10 and ICD-O3 do not have these two characteristics, we could not follow such a strategy and we had to propose an alternative one.

4.6.4 Integration process advantages

An analytical reading of our semantic integration process gives the possibility to understand and correct our methodological choices but also to indirectly observe limitations in the structure of SNOMED CT, ICD-10 and ICD-O3. Indeed, the anchoring and derivation stages are based on the SNOMED CT structure and, in particular, on the subsumption relations existing within SNOMED CT. Our method was able to identify some limitations and specificities of the SNOMED CT structure already described in the literature, such as the *"absence of difference in the description between children and parents"* [179, 210]. For example, SNOMED CT does not consider 109271004-*Melanocytic nevus of lip (disorder)* as being a benign tumor, erroneously. Although being correct, its anchor with the ICD-10 concept D22.0-*Melanocytic naevi of lip* was deleted during the filtering step (by tumor behavior) because it corresponded to a confounding conflict between the two concepts. Another example is the anchor of the ICD-O3 morphology concept 8151/3-*Insulinoma, malignant* with the SNOMED CT concept 20955008-*Insulinoma, malignant (morphologic abnormality)*. Although this anchor is correct, it was erroneously derived with some ICD-10 concepts of benign tumors because the SNOMED CT concept 20955008 is a descendant of 3898006-*Neoplasm, benign (morphologic abnormality)* and 86049000-*Malignant neoplasm, primary (morphologic abnormality)*. This inconsistency illustrates an uncontrolled use of the subsumption relationship in SNOMED CT, which is called *is_a* overloading [179].

Other knowledge resource-related problems were encountered during the semantic integration process. Indeed, from the beginning of the process, we identified concepts that did not participate in anchors. The concepts which could not be mapped are mainly ICD-O3 topography concepts with codes (.8) describing an overlapping anatomical site (*e.g.*, C63.8-*Overlapping lesion of male genital organs*), as well as ICD-10 and ICD-O3 concepts which use the category "other" for unlisted diagnoses or histological lesions. The ICD-10 concept C45.7-*Mesothelioma of other sites* is such an example. ICD-10 enumerates three anatomical sites for mesothelioma (C45.0-Mesothelioma of pleura, C45.1-*Mesothelioma of peritoneum* and C45.2-*Mesothelioma of pericardium*), and C45.7 encodes for all mesothelioma that are not pleura, peritoneum and pericardium mesothelioma [25, 26]. This representation is made because of the epidemiologic objectives of ICD-10 and ICD-O3. The objectives of SNOMED CT being different, it does not include such concepts. To address this issue, we could look for structural proximities between the concepts belonging to the "other" category and concepts already anchored, like ServOMap [18] and SAMBO [111] do. For these particular concepts, we could indeed search for their parent concepts having anchors with a SNOMED CT concept and some of the direct descendants of this SNOMED CT concept, which are not already anchored, could be mapped to the ICD-10 / ICD-O3 concept belonging to the "other" category.

4.6.5 Focus on semantic conflicts

In conclusion, we presented in this chapter an **automatic detection and correction of naming, scaling and confounding conflicts** based on the similarities established between their interpretants (D) (contexts (Co) and formal definitions (Df)) followed by a **resolution of open conflicts** using an automatic semantic approach. This approach is punctuated by manual interventions, in particular for the similarities between the contexts during the filtering step and for the detection of the appropriate transversal relations in SNOMED CT.

For a more automatic process taking into account all the lessons learned from the previous implemented processes, we present in the next chapter a generic process that can automatically take into account the specificity of any kind of knowledge resources and then perform an automatic mapping process that can overcome any type of semantic conflicts.

Any fool can know. The point is to understand.

Albert Einstein

Chapter 5

Conclusions and perspectives

In this chapter, we summarize the global findings presented in this document and the general perspectives induced by the implemented processes. The short-term perspectives of each implementation have been previously presented at the end of each chapter.

Our work addressed two major aspects in using support knowledge resources:

— a **semantic aspect** by allowing the resolution of different types of semantic conflicts. Indeed, SNOMED CT was used as a support for the integration of ICD-10 and ICD-O3 (chapter 4).

— a **practical aspect** by allowing the sharing of information across different information systems. This was the usage objective in the alignment of TLAB and LOINC (chapter 2), as well as the integration of RxNorm and SNOMED CT (chapter 3). Indeed, it was not LOINC and SNOMED CT themselves but their models that were used as supports to fill the gap respectively between TLAB and LOINC, and RxNorm and SNOMED CT on the other hand.

We firstly present the main outcomes of our work and the recommendations induced by our implementations. We then introduce the research lines that we envisage to explore in the coming years.

5.1 Findings

In this work, we detected and sometimes corrected semantic conflicts in the frame of different applications:

— **finding equivalent mappings** within the alignment process of TLAB and LOINC,

- **organizing the entities** of RxNorm and SNOMED CT into a coherent way thanks to an integration process based on their definitional features,
- **linking distinct but complementary entities** of ICD-10 and ICD-O3 in a semantically-enriched integration process.

From the analysis of techniques described in the literature and that we implemented, we found two emerging aspects:

- comparing the models of knowledge resources to be integrated as the first step of the alignment or integration processes is a helpful strategy and can be seen as a "best practice rule" to improve the existing mapping efforts.
- using background knowledge seems to be the most appropriate solution to detect and correct semantic conflicts. Specifically, this background knowledge can be a model (as shown within the alignment of TLAB and LOINC, and within the integration of RxNorm and SNOMED CT) or a knowledge resource (as illustrated within the integration of ICD-10 and ICD-O3).

While in the literature and/or our implementations, such background knowledge may be a domain expert, a model or an ontology, we believe that support ontologies constitute the best option for achieving a semantically-enriched integration process. In practice, such a support ontology must provide:

- synonyms for entities of knowledge resources to be linked,
- relationships between distinct hierarchies,
- hierarchy and formal definitions for entities.

These requirements imply that:

- the support ontology must cover the knowledge domains of resources to be integrated,
- the structure of the support ontology must not contain errors.

In practice, it is difficult to find ontologies with a structure free from any inconsistencies. However, the enhancement of the integration process we propose gives the possibility to find limitations or errors in the structure of the resources to be integrated and of the support ontology itself. Indeed, when using SNOMED CT as a support for integrating ICD-10 and ICD-O3, their structures have indirectly been audited [35]. Thus, using an ontology as a support in an integration process can help to identify inconsistencies in all the involved knowledge resources.

5.2 Recommendations

According to the remaining semantic conflicts observed during the alignment and integration processes we implemented (chapters 2 and 3), and the proposed strategy based on the framework of Alekovsky *et al.* [30] to semantically enrich the integration process (chapter 4), we propose a general integration process that consists in the two stages of anchoring and derivation as follows:

- **the anchoring.** This stage consists in finding equivalence relations between the entities of the resources to be integrated and those of the support ontology. It is composed of the five following steps:
  1. representing the entities using the <R,T,D> triplet,
  2. computing lexical and/or structural similarities between the entities of the resources to be integrated and those of the support ontology,
  3. eliminating mappings between entities that belong to disjoint contexts (Co),
  4. disambiguating multiple mappings using the structure of the support ontology, *i.e.*, choosing the entities of the support ontology that exhibit the more general formal definition (Df),
  5. stating equivalences between the entities of the resources to be integrated and those of the support ontology.
- **the derivation.** This stage consists in finding the appropriate relations between the anchored entities. It comprises three steps:
  1. identifying indirect equivalence or hierarchical relations between the entities of the knowledge resources,
  2. identifying the relations relating the context of the entities (Co) (hierarchical context) in the support ontology (*e.g.*, manually or via SPARQL Protocol and RDF Query Language (SPARQL [1]) queries),
  3. inferring the relation (through queries over logical definitions (Df)) between the entities related through their contexts. These queries must be realized from the most precise to the most general transversal relations that have been found.

5.3 Challenges and opportunities

Firstly, performing a semantic enrichment of the integration process gave the possibility to indirectly improve the quality of knowledge resources. One

[1] https://www.w3.org/TR/rdf-sparql-query/

typical error, known as "*is_a* overloading", has been encountered when performing the semantically-enriched integration process. Already identified in SNOMED CT [179] and in the NCI thesaurus [211], the "*is_a* overloading" results from erroneous heritage described within defined concepts. Thus, in our framework, cases of "*is_a* overloading" correspond to contradictions between the hierarchy (Co) and the formal definitions (Df) [35] and/or textual definitions (Dn).

Secondly, the solution we proposed is obviously limited by the possibility to find the appropriate support ontology. Some authors attempted to automatically find such support ontologies [185] by using online ontology search engines. Others proposed a strategy to perform the selection of the support ontology in an arbitrary set of ontologies [212, 213]. However, as in the processes we implemented, the selection of a support ontology or its discovery among an arbitrary set of ontologies is still a key issue. Indeed, it remains an intuitive task for experts in knowledge engineering to select the support ontology or to constitute the appropriate pool of candidate support ontologies. Nonetheless, we can notice that some knowledge resources are becoming references in their domain (*e.g.*, the FMA for human anatomy, the Gene Ontology for gene functions). In these cases, the use of these resources as a support should be favoured in their corresponding domains.

Finally, some limitations in the use of a support ontology can be mentioned. Sometimes, the detection and correction of certain semantic conflicts need some information that is not described in the structure of the support ontology. For example, confounding conflicts can also be consecutive to the evolution of knowledge and consequently related to the versioning of knowledge resources. In this case, an external model that manages the versioning of the support ontology would be necessary to find and resolve such conflicts. Thus, it could highlight a lack of coherence in mappings between entities from a version to another one. In addition, if no existing ontology contains the appropriate transversal relation between two complementary concepts, such link can be sought using background knowledge other than ontologies [214, 215] (*e.g.*, using scientific articles to find the appropriate relation between genes and diseases and standardize them into a new artefact). This strategy can also be used to enrich the description of knowledge resources.

5.4 Perspectives

5.4.1 Strategies to be explored

The use of machine learning techniques

In our implemented processes, we mainly used traditional lexical (non-vectorial) methods to perform natural language processing. However, as previously noticed in chapter 2, machine learning and deep learning are taking more and more place in the strategies of natural language processing [216–218]. The application of machine learning algorithms using as features some entities from knowledge resources that make use of non-vectorial methods (like the UMLS) is a strategy recently used in the medical domain [219–221].

To obtain better similarities between labels (L), the integration of a machine learning step in the processes of the creation of mappings is a perspective to be explored in future work. For example, in section 2.4.3, the lexical mapping of tokens followed by the validation of mappings between labels sharing the highest number of tokens in common could be replaced by machine learning algorithms with the expectation of better results.

The use of multiple ontologies as a support

When a knowledge resource covers multiple domains, it is more likely to be less accurate than a resource created specifically for the domains it should cover. As described in section 4.6.2, when the support ontology does not fully cover a specific domain, the use of another support for allowing a better anchoring step must be considered. The additional ontology must be specific to the anchored domain. In future works, we thus intend to study the benefit of using a specific support if a poor coverage is obtained during the anchoring step. In particular, to compensate the specific and questionable way in which SNOMED CT represents anatomy, we plan to use the FMA as a support ontology to improve the integration of ICD-O3 topography and SNOMED CT anatomy. We expect that the use of such additional ontology results in a better derivation.

The use of textual definitions

Our work was mainly based on similarities between labels (L) and interpretants (D). For interpretants, the characteristics we used are: the contexts (Co), the formal definitions (Df) and the post-coordinated expressions (Dc). We did not use textual definitions (Dn) available in knowledge resources. Thus, this can be another subject of future works.

Indeed, existing techniques explored the creation of formal definitions (Df) from textual definitions (Dn). Proposed techniques, like in [222, 223], depend on the source of textual definitions or the learning algorithm that may affect the quality of the resulting formal definitions. Confronting such a created formal definition permits to check its quality, which means going through an *auditing process*. Furthermore, for entities described by different types of definiens (Df, Dn, Dc, and Co), each type of interpretant must not give a contradictory definition. If they do not correspond exactly to the same definition, they must at least provide complementary definitions. Confronting the intended definition contained in the textual definition (Dn) to the stated formal definition (Df) (or a formal definition (Df) provided by an external knowledge resource) can allow the auditing process of knowledge resources but also improve the quality of the anchoring step.

5.4.2 Operationalization of the <R,T,D> triplet

We used the <R,T,D> triplet to have a solid scientific background to explore our implemented works. Nevertheless, through all the implemented processes, we manually identified the context (Co) and other descriptions of each entity in the manipulated knowledge resources. If the triplet influences properly our way of thinking and exploring the mapping results, the knowledge entities were not automatically annotated using the triplet. Thus, to facilitate future work using our methodological approach, it is important to perform the operationalization of the triplet and the standardization of the use of support ontologies to semantically enrich the integration process.

The operationalization of the <R,T,D> triplet is under development. To this end, we would like to use a specific standard developed in the ERIAS (Equipe de Recherche en Informatique Appliquée en Santé) team, called K-WARE (Knowledge warehouse) [224]. K-WARE is a meta-model that integrates all the formats used in the representation of knowledge resources (*e.g.*, RDF, SKOS, OWL, CSV). By integrating knowledge resources in K-WARE, the problematic of translation (or morphing) is thus already taken into account. Based on K-WARE, the steps defined for the anchoring and the derivation stages could be automatized. The first step of anchoring can be realized by an automatized annotation of knowledge resources using the <R,T,D> triplet formalism. The second step can be realized through the use of tools implementing morphosyntactic and/or structural techniques (*e.g.*, ServoMap). The other steps can be achieved by automatically taking into account the triplet structure of each entity and the mapping obtained beforehand. Thus, we hope to build an enriched K-WARE that is able to participate in international scientific campaigns such as OAEI to test the effectiveness and the generalizability of our findings. Participating in such campaigns can also help to improve the proposed process.

5.4.3 The integration of omics data

In addition to the generalization of the semantically-enriched integration process, we are also interested in the integration of omics data in the frame of personalized medicine. The interpretation of gene roles in diseases, and action in medicinal products can be better described. As previously said, for now, the mapping of knowledge resources representing genes, biological pathways, medicinal products and diseases is limited to simple matrix tables or specific mapping relations.

We plan to explore existing mappings, like those available in DisGeNET [225] or MalaCards [226], to semantically enrich the integration process they implemented. DisGenet is an open-access database that describes gene-disease associations using a specific proximity score (which is computed according to their simultaneous occurrence in PubMed articles) and the Semanticscience Integrated Ontology (SIO)[2] [227]. SIO is an upper-level ontology describing main associations between genes and diseases through 15 relations that are organized hierarchically. In MalaCards, the gene-disease (and/or disease-drugs) associations are represented as simple matrix tables [226].

Thus, the various omics databases differ in the precision of their mapping relations and covered scope. Consequently, they can be used to perform a semantically-enriched integration process (each bringing their mapping relations and scopes). A preliminary step will be a review of all the characteristics of available omics databases. Then, based on these characteristics, we will be able to describe a process for (1) the refinement of the relation between genes and diseases, and (2) the semantically-enriched integration with biological pathways and medicinal products. Thereafter, when many genes (or even biological pathways) are linked to a disease (or a medicinal product), tools like GSAn [228] can be used to choose the more appropriate set of genes (or set of biological pathways) for explaining these diseases (or medicinal products).

[2] https://raw.githubusercontent.com/micheldumontier/semanticscience/master/ontology/sio/release/sio-release.owl

Bibliography

1. Joubert, M., Abdoune, H., Merabti, T., Darmoni, S. & Fieschi, M. *Assisting the Translation of SNOMED CT into French using UMLS and four Representative French-language Terminologies* in *AMIA Annual Symposium Proceedings* (San Francisco, CA, USA, 2009), 291–295.
2. Studer, R., Benjamins, V. R. & Fensel, D. Knowledge engineering: Principles and methods. en. *Data & Knowledge Engineering* **25,** 161–197. ISSN: 0169023X (Mar. 1998).
3. Bourigault, D., Aussenac-Gilles, N. & Charlet, J. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle* **18,** 87–110 (2004).
4. Merabti, T., Abdouneb, H., Lecroq, T., Joubert, M. & Darmoni, S. J. Projection des relations SNOMED CT entre les termes de deux terminologies (CIM10 et SNOMED 3.5). *Risques, Technologies de l'Information pour les Pratiques Médicales*, 79–88 (2009).
5. Meystre, S. M. *et al.* Clinical data reuse or secondary use: current status and potential future progress. *Yearbook of medical informatics* **26,** 38–52 (2017).
6. Vogenberg, F. R., Barash, C. I. & Pursel, M. Personalized medicine: part 1: evolution and development into theranostics. *Pharmacy and Therapeutics* **35,** 560–567 (2010).
7. Garcia, I., Kuska, R. & Somerman, M. Expanding the Foundation for Personalized Medicine: Implications and Challenges for Dentistry. *Journal of Dental Research* **92,** S3–S10 (2013).
8. Sheth, A. P. in *Interoperating geographic information systems* 5–29 (Springer, 1999).
9. Da Silva, C. F. *et al.* Semantic Interoperability of Heterogeneous Semantic Resources. en. *Electronic Notes in Theoretical Computer Science* **150,** 71–85. ISSN: 15710661 (Mar. 2006).

10. Euzenat, J., Meilicke, C., Stuckenschmidt, H., Shvaiko, P. & Trojahn, C. in *Journal on Data Semantics XV* 158–192 (Springer, 2011).

11. Kalfoglou, Y. & Schorlemmer, M. IF-Map: An ontology-mapping method based on information-flow theory. *Journal on Data Semantics* **1,** 98–127 (2003).

12. Pinto, H. S., Gómez-Pérez, A. & Martins, J. P. Some Issues on Ontology Integration. *In Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends* **18** (Aug. 1999).

13. Fahad, M., Moalla, N., Bouras, A., Qadir, M. A. & Farukh, M. *Disjoint-knowledge analysis and preservation in ontology merging process* in *2010 Fifth International Conference on Software Engineering Advances* (Nice, France, 2010), 422–428.

14. McDonald, C. J. *et al.* LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry* **49,** 624–633 (2003).

15. Bodenreider, O., Cornet, R. & Vreeman, D. J. Recent Developments in Clinical Terminologies—SNOMED CT, LOINC, and RxNorm. *Yearbook of medical informatics* **27,** 129–139 (2018).

16. Nikiema, J. N., Mougin, F. & Jouhet, V. *Processus de prétraitement des libellés d'une terminologie d'interface* in *4e édition du Symposium sur l'Ingénierie de l'Information Médicale* (2017), 95–103.

17. Mary, M., Soualmia, L. F. & Gansel, X. *Formalisation de la terminologie LOINC et évaluation de ses avantages pour la classification des tests de laboratoire* in *28es Journées francophones d'Ingénierie des Connaissances IC 2017* (2017), 2–13.

18. Diallo, G. An effective method of large scale ontology matching. eng. *Journal of Biomedical Semantics* **5,** 44. ISSN: 2041-1480 (2014).

19. Bodenreider, O. & James, J. *The New SNOMED CT International Medicinal Product Model* en. in *Proceedings of the International Conference on Biological Ontology (ICBO 2018)* (Oregon, USA, June 2018).

20. European Medicines Agency, E. M. A. *Introduction to ISO Identification of Medicinal Products, SPOR programme* Nov. 2016. <https://www.ema.europa.eu/documents/other/introduction-iso-identification-medicinal-products-spor-programme%5C_en.pdf> (visited on 01/12/2018).

21. Abburu, S. A survey on ontology reasoners and comparison. *International Journal of Computer Applications* **57,** 33–39 (Nov. 2012).

22. Kazakov, Y., Krötzsch, M. & Simančík, F. *ELK Reasoner: architecture and Evaluation* in *Proceedings of the OWL Reasoner Evaluation Workshop 2012* (CEUR Workshop Proceedings, July 2012).

23. Dentler, K. & Cornet, R. Intra-axiom redundancies in SNOMED CT. eng. *Artificial Intelligence in Medicine* **65,** 29–34. ISSN: 1873-2860 (Sept. 2015).

24. Nikiema, J. N. *Integrating RxNorm with medicinal products in SNOMED CT* Presentation at NLM/NIH, Bethesda. NIH, Bethesda,Maryland, 2018. <https://mor.nlm.nih.gov/pubs/alum/2018-nikiema-pres.pdf>.

25. World Health Organization, W. H. O. *International Statistical Classification of Diseases and Related Health Problems 10th revision* 2010th ed. ISBN: 978 92 4 154834 2. <http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf> (visited on 04/10/2017) (2011).

26. Fritz, A. *et al. International classification of diseases for oncology: ICD-O* 3rd ed. eng (ed Fritz, A.) OCLC: 248314653. ISBN: 978-92-4-154534-1 (World Health Organization, Geneva, 2000).

27. Thieblin, E., Haemmerlé, O., Hernandez, N. & Trojahn, C. Survey on complex ontology matching. *semantic web journal* (2019).

28. Ngo, D., Bellahsene, Z. & Todorov, K. *Opening the black box of ontology matching* in *Extended Semantic Web Conference* (2013), 16–30.

29. Tigrine, A. N., Bellahsene, Z. & Todorov, K. *Light-weight cross-lingual ontology matching with LYAM++* in *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (2015), 527–544.

30. Aleksovski, Z., Klein, M., ten Kate, W. & van Harmelen, F. in *Managing Knowledge in a World of Networks* (eds Staab, S. & Svátek, V.)red. by Hutchison, D. *et al.*, 182–197 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006). doi:10.1007/11891451_18.

31. Shvaiko, P. & Euzenat, J. Ontology matching: state of the art and future challenges. *IEEE Transactions on knowledge and data engineering* **25,** 158–176 (2013).

32. Dhombres, F. & Bodenreider, O. Interoperability between phenotypes in research and healthcare terminologies—Investigating partial mappings between HPO and SNOMED CT. *Journal of Biomedical Semantics* **7,** 3 (2016).

33. Mougin, F., Dupuch, M. & Grabar, N. *Improving the mapping between MedDRA and SNOMED CT* in *Artificial Intelligence in Medicine* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), 220–224. ISBN: 978-3-642-22218-4.

34. Kim, T. Y., Coenen, A. & Hardiker, N. Semantic mappings and locality of nursing diagnostic concepts in UMLS. en. *Journal of Biomedical Informatics* **45,** 93–100. ISSN: 15320464 (Feb. 2012).

35. Nikiema, J. N., Jouhet, V. & Mougin, F. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. *Journal of Biomedical Informatics* **74,** 46–58 (2017).

36. Jouhet, V., Mougin, F., Bréchat, B. & Thiessard, F. Building a model for disease classification integration in oncology, an approach based on the National Cancer Institute thesaurus. *Journal of Biomedical Semantics* **8,** 6 (2017).

37. Pesquita, C., Faria, D., Santos, E. & Couto, F. M. *To repair or not to repair: reconciling correctness and coherence in ontology reference alignments* in *Proceedings of the 8th International Conference on Ontology Matching-Volume 1111* (CEUR-WS. org, 2013), 13–24.

38. Zhu, X., Fan, J.-W., Baorto, D. M., Weng, C. & Cimino, J. J. A review of auditing methods applied to the content of controlled biomedical terminologies. *Journal of Biomedical Informatics* **42,** 413–425 (2009).

39. Zweigenbaum, P. Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé* **2,** 5 (1999).

40. Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L. & Brown, S. H. A model for evaluating interface terminologies. *Journal of the American Medical Informatics Association* **15,** 65–76 (2008).

41. Merabti, T., Soualmia, L. F., Grosjean, J., Joubert, M. & Darmoni, S. J. *Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications* in *Medical Informatics* (ed Mordechai, S.) (InTech, Rijeka, 2012). Chap. 3. doi:10.5772/37738.

42. IHTSDO. *SNOMED CT Technical implementation Guide January 2015 International Release (GB English)* 2015. <http://ihtsdo.org/fileadmin/user%5C_upload/doc/download/doc%5C_TechnicalImplementationGuide%5C_Current-en-GB%5C_INT%5C_20150131.pdf?ok> (visited on 04/18/2016).

43. Deckard, J., McDonald, C. J. & Vreeman, D. J. Supporting interoperability of genetic data with LOINC. *Journal of the American Medical Informatics Association* **22,** 621–627 (2015).

44. Wang, Y., Patrick, J., Miller, G. & O'Hallaran, J. A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Medical Informatics and Decision Making* **8,** S5. ISSN: 1472-6947 (2008).

45. De Keizer, N. F., Abu-Hanna, A., Zwetsloot-Schonk, J., *et al.* Understanding terminological systems I: terminology and typology. *Methods of Information in Medicine* **39,** 16–21 (2000).

46. Roche, C. *Ontoterminology: How to unify terminology and ontology into a single paradigm* in *LREC 2012, Eighth International Conference on Language Resources and Evaluation* (2012), 2626–2630.

47. Bodenreider, O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics,* 67–79 (2008).

48. Luciano, J. S. *et al.* The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. *Journal of Biomedical Semantics* **2,** S1 (2011).

49. Zinsstag, J., Schelling, E., Waltner-Toews, D. & Tanner, M. From "one medicine" to "one health" and systemic approaches to health and well-being. *Preventive veterinary medicine* **101,** 148–156 (2011).

50. Safran, C. *et al.* Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. en. *Journal of the American Medical Informatics Association* **14,** 1–9. ISSN: 1067-5027, 1527-974X (Jan. 2007).

51. Murdoch, T. B. & Detsky, A. S. The inevitable application of big data to health care. *JAMA* **309,** 1351–1352 (2013).

52. MacKay, E. N. & Sellers, A. H. The Ontario cancer incidence survey, 1964-1966: a new approach to cancer data acquisition. eng. *Canadian Medical Association Journal* **109,** 489. ISSN: 0008-4409 (Sept. 1973).

53. *Cancer registration: principles and methods* eng (eds Jensen, O. M., Parkin, D. M., International Association of Cancer Registries & International Agency for Research on Cancer) *IARC scientific publications* **95.** OCLC: 24293858. ISBN: 978-92-832-1195-2 (Internat. Agency for Research on Cancer [u.a.], Lyon, 1991).

54. Parkin, D. M. The role of cancer registries in cancer control. en. *International Journal of Clinical Oncology* **13,** 102–111. ISSN: 1341-9625, 1437-7772 (Apr. 2008).

55. Jouhet, V., Defossez, G., CRISAP, CoRIM & Ingrand, P. Automated Selection of Relevant Information for Notification of Incident Cancer Cases within a Multisource Cancer Registry: en. *Methods of Information in Medicine* **52,** 411–421. ISSN: 0026-1270 (Apr. 2013).

56. Colin, C. *et al.* Data Quality in a DRG-Based Information System. en. *International Journal for Quality in Health Care* **6,** 275–280. ISSN: 1353-4505, 1464-3677 (Sept. 1994).

57. Jain, K. K. Personalized medicine. *Current opinion in molecular therapeutics* **4,** 548–558. ISSN: 1464-8431 (Dec. 2002).

58. Cho, S.-H., Jeon, J. & Kim, S. I. Personalized medicine in breast cancer: a systematic review. *Journal of breast cancer* **15,** 265–272 (2012).

59. Sugaya, N. *et al.* An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. *BMC pharmacology* **7,** 10 (2007).

60. Mao, M. & Peng, Y. *Ontology mapping: towards semantic interoperability in distributed and heterogeneous environments* eng. OCLC: 837452301. ISBN: 978-3-8383-4229-0 (Lambert Acad. Publ, Saarbrücken, 2010).

61. Singh, M., Jain, S. & Panchal, V. *Modeling semantic Heterogeneity in Dataspace: A Machine Learning Approach* in *Information Technology (ICIT), 2014 International Conference on* (IEEE, Dec. 2014), 275–280. doi:10.1109/ICIT.2014.24.

62. Lesnikova, T. *RDF Data Interlinking: evaluation of Cross-lingual Methods* PhD thesis (Université Grenoble Alpes, 2016). <https://tel.archives-ouvertes.fr/tel-01366030/> (visited on 01/05/2017).

63. Bouquet, P. *et al. D2.2.1 Specification of a common framework for characterizing alignment* 2004.

64. Chalupsky, H. *Ontomorph: A translation system for symbolic knowledge* in *KR* (2000), 471–482.

65. Bizid, I., Faiz, S., Boursier, P. & Yusuf, J. C. M. *Integration of heterogeneous spatial databases for disaster management* in *International Conference on Conceptual Modeling* (2013), 77–86.

66. Kaladevi, R. & Mirnalinee, T. T. Heterogeneous information management using ontology mapping. *ARPN Journal of Engineering and Applied Sciences* **10,** 2078–2081 (Jan. 2015).

67. Corcho, O. *et al. Evaluation experiment of ontology tools' interoperability with the WebODE ontology engineering workbench* in *Proceedings of the 2nd International Workshop on Evaluation of Ontology-based Tools held at the 2nd International Semantic Web Conference ISWC 2003* **87.** Ontology Engineering Group (CEUR, Oct. 2003).

68. Noy, N. F. *et al.* BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research* **37,** W170–W173 (2009).

69. Seremeti, L. & Kameas, A. in *Theory and Applications of Ontology: Computer Applications* 131–154 (Springer, Dordrecht, 2010). doi:https://doi.org/10.1007/978-90-481-8847-5_6.

70. Giannangelo, K. & Millar, J. Mapping SNOMED CT to ICD-10. eng. *Studies in Health Technology and Informatics* **180,** 83–87. ISSN: 0926-9630 (2012).

71. Souvignet, J. & Rodrigues, J.-M. Toward a Patient Safety Upper Level Ontology. *Studies in Health Technology and Informatics,* 160–164. ISSN: 0926-9630 (2015).

72. Fung, K. W., Xu, J., Ameye, F., Gutiérrez, A. R. & D'Havé, A. *Leveraging lexical matching and ontological alignment to map SNOMED CT surgical procedures to ICD-10-PCS* in *AMIA Annual Symposium proceedings* **2016** (2016), 570–579.

73. Kolyvakis, P., Kalousis, A., Smith, B. & Kiritsis, D. Biomedical ontology alignment: an approach based on representation learning. *Journal of biomedical semantics* **9,** 21 (2018).

74. Rodrigues, J. M. *et al. Semantic Alignment between ICD-11 and SNOMED CT.* in *MedInfo* (2015), 790–794.

75. Harrow, I. *et al.* Matching disease and phenotype ontologies in the ontology alignment evaluation initiative. *Journal of biomedical semantics* **8,** 55 (2017).

76. Tulasi, R. L. & Rao, M. S. Survey on Techniques for Ontology Interoperability in Semantic Web. *Global Journal of Computer Science and Technology* **14,** 57–62 (2014).

77. Al-Baltah, I. A., Ghani, A., RAHMAN, W. & Atan, R. A comparative study on ontology development methodologies towards building semantic conflicts detection ontology for heterogeneous web services. *Research Journal of Applied Sciences, Engineering and Technology* **7,** 2674–2679 (2014).

78. Li, C. & Ling, T. W. *OWL-based semantic conflicts detection and resolution for data interoperability* in *International conference on conceptual modeling* (2004), 266–277.

79. Ram, S. & Park, J. Semantic Conflict Resolution Ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts. *IEEE Transactions on Knowledge and Data engineering* **16,** 189–202 (2004).

80. Schuster, G. & Stuckenschmidt, H. *Building shared ontologies for terminology integration* in *KI-01 Workshop on Ontologies, Vienna, Austria* (2001).

81. Bello, S. M. *et al.* Disease Ontology: improving and unifying disease annotations across species. *Disease models & mechanisms* **11,** dmm032839 (2018).

82. Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic acids research* **43,** D1071–D1078 (2014).

83. Finke, M. T., Filice, R. W. & Kahn Jr, C. E. Integrating ontologies of human diseases, phenotypes, and radiological diagnosis. *Journal of the American Medical Informatics Association* **26,** 149–154 (2019).

84. Filice, R. W. & Kahn, C. E. Integrating an Ontology of Radiology Differential Diagnosis with ICD-10-CM, RadLex, and SNOMED CT. *Journal of digital imaging,* 1–5 (2019).

85. Xiang, Y. & Janga, S. C. Building integrated ontological knowledge structures with efficient approximation algorithms. *BioMed research international* **2015** (2015).

86. Salamon, J., Reginato, C. & Barcellos, M. *Ontology Integration Approaches: A Systematic Mapping* in *Proceedings of the XI Seminar on Ontology Research in Brazil and II Doctoral and Masters Consortium on Ontologies* (São Paulo, Brazil, Oct. 2018), 161–172.

87. Xue, H., Peng, J. & Shang, X. Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. *BMC Systems Biology* **13,** 34 (2019).

88. Wache, H. & Stuckenschmidt, H. in *Modeling and Using Context* (eds Akman, V., Bouquet, P., Thomason, R. & Young, R.)red. by Goos, G., Hartmanis, J. & van Leeuwen, J., 367–380 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2001). doi:10.1007/3-540-44607-9_28.

89. Elasri, H., Sekkaki, A. & Kzaz, L. *An ontology-based method for semantic integration of business components* in *11th Annual International Conference on New Technologies of Distributed Systems (NOTERE)* (2011), 1–8.

90. Hajmoosaei, A. & Abdul-Kareem, S. *An ontology-based approach for resolving semantic schema conflicts in the extraction and integration of query-based information from heterogeneous web data sources* in *Proceedings of the Third Australasian Workshop on Advances in Ontologies - Volume 85* (2007), 35–43.

91. Goh, C. H. *Representing and reasoning about semantic conflicts in heterogeneous information systems* PhD thesis (Massachusetts Institute of Technology, 1996).

92. Wache, H. *et al. Ontology-based integration of information-a survey of existing approaches* in *IJCAI-01 workshop: ontologies and information sharing* **2001** (Citeseer, 2001), 108–117.

93. Jirkovský, V. & Ichise, R. in *Semantic Technology* (eds Kim, W., Ding, Y. & Kim, H.-G.) 348–363 (Springer International Publishing, Cham, 2014).

94. Peirce, C. S. *Collected papers of charles sanders peirce* (Harvard University Press, Boston, USA, 1974).

95. Raggatt, P. T. The dialogical self and thirdness: A semiotic approach to positioning using dialogical triads. *Theory & Psychology* **20,** 400–419 (2010).

96. Siegel, M. More than words: The generative power of transmediation for learning. *Canadian Journal of Education/Revue canadienne de l'éducation,* 455–475 (1995).

97. Gomes, A., Gudwin, R. & Queiroz, J. Towards meaning processes in computers from Peircean semiotics. *SEED Journal (Semiotics, Evolution, Energy, and Development)* **3,** 69–79 (2003).

98. Buczynskagarewicz, H. The interpretant and a system of signs. *Ars Semeiotica* **4,** 187–200 (1981).

99. Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P. & Nardi, D. *The description logic handbook: Theory, implementation and applications* (Cambridge university press, Cambridge , UK, 2003).

100. Chute, C. G. The rendering of human phenotype and rare diseases in ICD-11. *Journal of inherited metabolic disease* **41,** 563–569 (2018).

101. Lopetegui, M. & Mauro, A. A Novel Approach to Create a Machine Readable Concept Model for Validating SNOMED CT Concept Post-coordination. *Studies in health technology and informatics* **216,** 1087–1087 (2015).

102. Horridge, M. *et al. The Manchester OWL Syntax.* in *OWLED* **216** (2006).

103. Schulz, S., Kumar, A. & Bittner, T. Biomedical ontologies: What part-of is and isn't. *Journal of Biomedical Informatics* **39.** Biomedical Ontologies, 350–361. issn: 1532-0464 (2006).

104. Detwiler, L. T., Mejino, J. L. & Brinkley, J. F. From frames to OWL2: Converting the Foundational Model of Anatomy. *Artificial Intelligence in Medicine* **69,** 12–21. issn: 0933-3657 (2016).

105. Mejino Jr, J. L., Agoncillo, A. V., Rickard, K. L. & Rosse, C. *Representing complexity in part-whole relationships within the foundational model of anatomy* in *AMIA Annual Symposium Proceedings* **2003** (2003), 450.

106. Saitwal, H. *et al.* Cross-terminology mapping challenges: a demonstration using medication terminological systems. eng. *Journal of Biomedical Informatics* **45,** 613–625. ISSN: 1532-0480 (Aug. 2012).

107. Euzenat, J. & Shvaiko, P. *Ontology matching* (Springer Science & Business Media, 2013).

108. Htun, H. H. & Sornlertlamvanich, V. *Text similarity approach for SNOMED CT primitive concept similarity measure* in *8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)* (May 2017), 1–5. doi:10.1109/ICTEmSys.2017.7958766.

109. Schulz, S., Rodrigues, J.-M., Rector, A. & Chute, C. G. Interface Terminologies, Reference Terminologies and Aggregation Terminologies: A Strategy for Better Integration. *Studies in health technology and informatics* **245,** 940–944 (2017).

110. Bodenreider, O. & Zhang, S. *Comparing the representation of anatomy in the FMA and SNOMED CT* in *AMIA Annual Symposium Proceedings* **2006** (2006), 46.

111. Tan, H. & Lambrix, P. *SAMBO results for the ontology alignment evaluation initiative 2007* in *Proceedings of the 2nd International Conference on Ontology Matching-Volume 304* (CEUR-WS. org, Aachen, Germany, 2007), 236–243.

112. Tian, A., Sequeda, J. F. & Miranker, D. P. *QODI: Query as context in automatic data integration* in *International Semantic Web Conference* (Springer, 2013), 624–639. ISBN: 978-3-642-41335-3.

113. Cruz, I. F. *et al. Using AgreementMaker to align ontologies for OAEI 2011* in *Proceedings of the 6th International Conference on Ontology Matching-Volume 814* (CEUR-WS. org, 2011), 114–121.

114. Winnenburg, R. & Bodenreider, O. A framework for assessing the consistency of drug classes across sources. *Journal of Biomedical Semantics* **5,** 30 (2014).

115. Cruz, I. F. & Sunna, W. Structural alignment methods with applications to geospatial ontologies. *Transactions in GIS* **12,** 683–711 (2008).

116. Jiménez-Ruiz, E. & Grau, B. C. *Logmap: Logic-based and scalable ontology matching* in *International Semantic Web Conference* (2011), 273–288. ISBN: 978-3-642-25073-6.

117. Jiménez, E., Meilicke, C., Grau, B. C., Horrocks, I., *et al. Evaluating mapping repair systems with large biomedical ontologies* in *26th International Workshop on Description Logics* (2013).

118. Santos, E., Faria, D., Pesquita, C. & Couto, F. M. Ontology alignment repair through modularization and confidence-based heuristics. *PloS one* **10,** e0144807 (2015).

119. Dhombres, F., Winnenburg, R., Case James, T. & Bodenreider, O. Extending the coverage of phenotypes in SNOMED CT through post-coordination. *Studies in Health Technology and Informatics,* 795–799. ISSN: 0926-9630 (2015).

120. Dolin, R. H., Spackman, K. A. & Markwell, D. Selective retrieval of pre- and post-coordinated SNOMED concepts. eng. *AMIA Annual Symposium Proceedings,* 210–214. ISSN: 1531-605X (2002).

121. Sabou, M., d'Aquin, M. & Motta, E. Using the Semantic Web as Background Knowledge for Ontology Mapping. *Ontology Matching,* 1–12. ISSN: 1613-0073 (2006).

122. Ning, W., Yu, M. & Kong, D. Evaluating semantic similarity between Chinese biomedical terms through multiple ontologies with score normalization: An initial study. *Journal of Biomedical Informatics* **64,** 273–287 (2016).

123. Ji, X., Ritter, A. & Yen, P.-Y. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *Journal of Biomedical Informatics* **69,** 33–42 (2017).

124. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13,** R5 (2012).

125. Dragisic, Z. *et al. User validation in ontology alignment* in *The Semantic Web – ISWC 2016* (2016), 200–217. ISBN: 978-3-319-46523-4.

126. Miller, G. A. WordNet: a lexical database for English. *Communications of the ACM* **38,** 39–41 (1995).

127. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13,** 1 (2012).

128. Al-Baltah, I. A., Ghani, A. A. A., Ab Rahman, W. N. W. & Atan, R. Semantic conflicts detection of heterogeneous messages of web services: challenges and solution. *Journal of Computer Science* **10,** 1428 (2014).

129. Calero, J. M. A. *et al.* Detection of semantic conflicts in ontology and rule-based information systems. *Data & Knowledge Engineering* **69,** 1117–1137 (2010).

130. Liu, Q., Huang, T., Liu, S.-H. & Zhong, H. An ontology-based approach for semantic conflict resolution in database integration. *Journal of Computer Science and Technology* **22,** 218–227 (2007).

131. Ouksel, A. M. & Naiman, C. F. Coordinating context building in heterogeneous information systems. *Journal of Intelligent Information Systems* **3,** 151–183 (1994).

132. Visser, P. R., Jones, D. M., Bench-Capon, T. J. & Shave, M. J. *Assessing heterogeneity by classifying ontology mismatches* in *Proceedings of the FOIS* **46** (1998), 148–162. ISBN: 978-90-5199-399-8.

133. Rosenbloom, S. T., Miller, R. A., Johnson, K. B., Elkin, P. L. & Brown, S. H. Interface Terminologies: Facilitating Direct Entry of Clinical Data into Electronic Health Record Systems. *Journal of the American Medical Informatics Association* **13,** 277–288. ISSN: 1067-5027, 1527-974X (May 2006).

134. Juvé-Udina, M. E. What patients' problems do nurses e-chart? Longitudinal study to evaluate the usability of an interface terminology. *International Journal of Nursing Studies* **50,** 1698–1710. ISSN: 00207489 (Dec. 2013).

135. Christel, D. *et al.* Standards and Specifications in Pathology: Image Management, Report Management and Terminology. en. *Studies in Health Technology and Informatics,* 105–122. ISSN: 0926-9630 (2012).

136. Griffon, N. *Modélisation, création et évaluation de flux de terminologies et de terminologies d'interface : application à la production d'examens complémentaires de biologie et d'imagerie médicale.* Theses (Université de Rouen, Oct. 2013).

137. Rosenbloom, S. T. *et al.* Using SNOMED CT to Represent Two Interface Terminologies. *Journal of the American Medical Informatics Association* **16,** 81–88. ISSN: 1067-5027, 1527-974X (Jan. 2009).

138. Wade, G. & Rosenbloom, S. T. Experiences mapping a legacy interface terminology to SNOMED CT. *BMC Medical Informatics and Decision Making* **8,** S3. ISSN: 1472-6947 (2008).

139. Oluoch, T. *et al.* A structured approach to recording AIDS-defining illnesses in Kenya: A SNOMED CT based solution. *Journal of Biomedical Informatics* **56,** 387–394. ISSN: 15320464 (Aug. 2015).

140. Griffon, N., Savoye-Collet, C., Massari, P., Daniel, C. & Darmoni, S. J. An interface terminology for medical imaging ordering purposes. *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2012,** 1237–1243. ISSN: 1942-597X (2012).

141. Bakhshi-Raiez, F., Ahmadian, L., Cornet, R., de Jonge, E. & de Keizer, N. Construction of an interface terminology on SNOMED CT. Generic approach and its application in intensive care. *Methods of Information in Medicine* **49,** 349 (2010).

142. Parr, S. K., Shotwell, M. S., Jeffery, A. D., Lasko, T. A. & Matheny, M. E. Automated mapping of laboratory tests to LOINC codes using noisy labels in a national electronic health record system database. en. *Journal of the American Medical Informatics Association* **25,** 1292–1300. issn: 1067-5027, 1527-974X (Oct. 2018).

143. Kopanitsa, G. Mapping Russian Laboratory Terms to LOINC. *Studies in health technology and informatics* **210,** 379–383 (2015).

144. Zunner, C., Bürkle, T., Prokosch, H.-U. & Ganslandt, T. Mapping local laboratory interface terms to LOINC at a German university hospital using RELMA V. 5: a semi-automated approach. *Journal of the American Medical Informatics Association* **20,** 293–297 (2013).

145. Lau, L. M., Johnson, K., Monson, K., Lam, S. H. & Huff, S. M. A method for the automated mapping of laboratory results to LOINC. eng. *Proceedings. AMIA Symposium,* 472–476. issn: 1531-605X (2000).

146. Baker, T. *et al.* Key choices in the design of Simple Knowledge Organization System (SKOS). en. *Journal of Web Semantics* **20,** 35–49. issn: 15708268 (May 2013).

147. Forrey, A. W. *et al.* Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry* **42,** 81–90 (1996).

148. McDonald, C. *et al. LOINC Users' Guide (June 2018)* en-US. June 2018. <`https://loinc.org/file-access/`> (visited on 12/10/2018).

149. Regenstrief Institute, T. & Société Française d'Informatique de Laboratoire. *LOINC translation users-guide* en-US. June 2010.

150. Baorto, D. M., Cimino, J. J., Parvin, C. A. & Kahn, M. G. Combining laboratory data sets from multiple institutions using the logical observation identifier names and codes (LOINC). en. *International Journal of Medical Informatics* **51,** 29–37. issn: 13865056 (July 1998).

151. Khan, A. N. *et al.* Standardizing Laboratory Data by Mapping to LOINC. en. *Journal of the American Medical Informatics Association* **13,** 353–355. issn: 1067-5027, 1527-974X (May 2006).

152. Kim, H., El-Kareh, R., Goel, A., Vineet, F. & Chapman, W. W. An approach to improve LOINC mapping through augmentation of local test names. en. *Journal of Biomedical Informatics* **45,** 651–657. ISSN: 15320464 (Aug. 2012).

153. Lee, L.-H., Groß, A., Hartung, M., Liou, D.-M. & Rahm, E. A multi-part matching strategy for mapping LOINC with laboratory terminologies. en. *Journal of the American Medical Informatics Association* **21,** 792–800. ISSN: 1067-5027 (2014).

154. Kopanitsa, G. Application of a Regenstrief RELMA V. 6.6 to Map Russian Laboratory Terms to LOINC. *Methods of information in medicine* **55,** 177–181 (2016).

155. Kukich, K. Technique for automatically correcting words in text. *ACM Computing Surveys* **24,** 377–439. ISSN: 03600300 (Dec. 1992).

156. Sproat, R. *et al.* Normalization of non-standard words. *Computer Speech & Language* **15,** 287–333. ISSN: 08852308 (July 2001).

157. Gadde, P., Subramaniam, L. & Faruquie, T. A. *Adapting a WSJ trained part-of-speech tagger to noisy text: preliminary results* in *Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data* (2011), 5.

158. Gospodnetić, O. & Hatcher, E. *Lucene in action* (Manning Publications, 2005).

159. Miles, A., Matthews, B., Wilson, M. & Brickley, D. *SKOS core: simple knowledge organisation for the web* in *International Conference on Dublin Core and Metadata Applications Proceedings* (2005), 3–10.

160. Srinivasan, A., Kunapareddy, N., Mirhaji, P. & Casscells, S. W. *Semantic web representation of LOINC: an ontological perspective* in *AMIA Annual Symposium Proceedings* (Washington, DC, USA, 2006), 1107.

161. Fidahussein, M. & Vreeman, D. J. A corpus-based approach for automated LOINC mapping. en. *Journal of the American Medical Informatics Association* **21,** 64–72. ISSN: 1067-5027 (May 2014).

162. Balahur, A. & Turchi, M. *Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data.* in *RANLP* (eds Angelova, G., Bontcheva, K. & Mitkov, R.) (RANLP 2013 Organising Committee / ACL, 2013), 49–55.

163. Oana-Sorina, L., Ciprian-Bogdan, C. & Lăcrămioara, S.-T. Harnessing Ontologies to Improve Prescription in Pediatric Medicine. *Studies in Health Technology and Informatics,* 97–101. ISSN: 0926-9630 (2018).

164. Souissi, S. B., Abed, M., Elhiki, L., Fortemps, P. & Pirlot, M. Reducing the Toxicity Risk in Antibiotic Prescriptions by Combining Ontologies with a Multiple Criteria Decision Model. eng. *AMIA ... Annual Symposium proceedings. AMIA Symposium* **2017,** 1625–1634. ISSN: 1942-597X (2017).

165. Shen, Y. *et al.* An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. en. *Artificial Intelligence in Medicine* **86,** 20–32. ISSN: 09333657 (Mar. 2018).

166. Zhang, Y. *et al.* Extracting drug-enzyme relation from literature as evidence for drug drug interaction. en. *Journal of Biomedical Semantics* **7.** ISSN: 2041-1480. doi:10.1186/s13326-016-0052-6 (Dec. 2016).

167. Tari, L., Anwar, S., Liang, S., Cai, J. & Baral, C. Discovering drug-drug interactions: a text-mining and reasoning approach based on properties of drug metabolism. en. *Bioinformatics* **26,** i547–i553. ISSN: 1367-4803, 1460-2059 (Sept. 2010).

168. Souvignet, J., Declerck, G., Asfari, H., Jaulent, M.-C. & Bousquet, C. OntoADR a semantic resource describing adverse drug reactions to support searching, coding, and information retrieval. en. *Journal of Biomedical Informatics* **63,** 100–107. ISSN: 15320464 (Oct. 2016).

169. Zhichkin, P. E., Athey, B. D., Avigan, M. I. & Abernethy, D. R. Needs for an Expanded Ontology-Based Classification of Adverse Drug Reactions and Related Mechanisms. *Clinical Pharmacology & Therapeutics* **91,** 963–965. ISSN: 0009-9236, 1532-6535 (June 2012).

170. Harpaz, R. *et al.* Performance of Pharmacovigilance Signal-Detection Algorithms for the FDA Adverse Event Reporting System. *Clinical Pharmacology & Therapeutics* **93,** 539–546. ISSN: 0009-9236, 1532-6535 (June 2013).

171. Lai, E. C. *et al.* Applying a common data model to Asian databases for multinational pharmacoepidemiologic studies: opportunities and challenges. en. *Clinical Epidemiology* **Volume 10,** 875–885. ISSN: 1179-1349 (July 2018).

172. Dhavle, A. A. *et al.* Evaluating the implementation of RxNorm in ambulatory electronic prescriptions. en. *Journal of the American Medical Informatics Association* **23,** e99–e107. ISSN: 1067-5027, 1527-974X (Apr. 2016).

173. Wang, L. *et al.* Toward a normalized clinical drug knowledge base in China—applying the RxNorm model to Chinese clinical drugs. en. *Journal of the American Medical Informatics Association* **25,** 809–818. ISSN: 1067-5027, 1527-974X (July 2018).

174. Hanna, J., Joseph, E., Brochhausen, M. & Hogan, W. R. Building a drug ontology based on RxNorm and other sources. en. *Journal of Biomedical Semantics* **4,** 44. ISSN: 2041-1480 (2013).

175. Nikiema, J. N. & Bodenreider, O. *Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability* en. in *Proceedings of the International Conference on Biological Ontology (ICBO 2019)* (Buffalo, New York, USA, Aug. 2019).

176. Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T. & Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. en. *Journal of the American Medical Informatics Association* **18,** 441–448. ISSN: 1067-5027, 1527-974X (July 2011).

177. Bodenreider, O. & Peters, L. B. A Graph-based Approach to Auditing RxNorm. *Journal of biomedical informatics* **42,** 558–570. ISSN: 1532-0464 (June 2009).

178. Hitzler, P., Gangemi, A. & Janowicz, K. *Ontology engineering with ontology design patterns: Foundations and applications* (IOS Press, 2016).

179. Bodenreider, O., Smith, B., Kumar, A. & Burgun, A. *Investigating subsumption in DL-based terminologies: A Case Study in SNOMED CT.* in *KR-MED* **2004** (2004), 12–20.

180. Schulz, S., Markó, K. & Suntisrivaraporn, B. Formal representation of complex SNOMED CT expressions. en. *BMC Medical Informatics and Decision Making* **8,** S9. ISSN: 1472-6947 (2008).

181. Cornet, R. & Schulz, S. Relationship groups in SNOMED CT. eng. *Studies in Health Technology and Informatics* **150,** 223–227. ISSN: 0926-9630 (2009).

182. Kazakov, Y., Krötzsch, M. & Simančík, F. *ELK Reasoner: Architecture and Evaluation.* in *Proceedings of the OWL Reasoner Evaluation Workshop 2012 (ORE'12)* **858** (CEUR-WS.org, 2012).

183. Pinkel, C. *et al. How to best find a partner? An evaluation of editing approaches to construct R2RML mappings* in *European Semantic Web Conference* (Springer, 2014), 675–690.

184. Safar, B., Reynaud, C. & Calvier, F. Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire. *1ières Journées Francophones sur les Ontologies (JFO 2007),* 21–35 (2007).

185. Sabou, M., D'Aquin, M. & Motta, E. in *Journal on data semantics XI* 156–190 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008). doi:10.1007/978-3-540-92148-6_6.

186. Gartina, I., Akbar, S., Sitohang, B. & Azizah, F. *Review of ontology matching with background knowledge* in *2016 International Conference on Data and Software Engineering (ICoDSE)* (Oct. 2016), 1–6. doi:10.1109/ICODSE.2016.7936159.

187. Zhang, S. & Bodenreider, O. Experience in Aligning AnatomicalOntologies. *International Journal on Semantic Web and Information Systems (IJSWIS)* **3,** 1–26 (2007).

188. Stuckenschmidt, H., Van Harmelen, F., Serafini, L., Bouquet, P. & Giunchiglia, F. Using C-OWL for the alignment and merging of medical ontologies. <http://eprints.biblio.unitn.it/523/> (visited on 11/22/2016) (2004).

189. Kim, T. Y., Coenen, A. & Hardiker, N. Semantic mappings and locality of nursing diagnostic concepts in UMLS. en. *Journal of Biomedical Informatics* **45,** 93–100. issn: 15320464 (Feb. 2012).

190. Winnenburg, R. *et al. Aligning Pharmacologic Classes Between MeSH and ATC.* in *Proceedings of the International Conference on Biological Ontology (ICBO)* (Montreal, Quebec, Canada, 2013).

191. Burgun, A. & Bodenreider, O. *Issues in integrating epidemiology and research information in oncology: Experience with ICD-O3 and the NCI Thesaurus* in *AMIA Annu Symp Proc* (2007), 85–89.

192. Jiang, G., Solbrig, H. R. & Chute, C. G. Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network. en. *Journal of Biomedical Informatics* **44,** S78–S85. issn: 15320464 (Dec. 2011).

193. Jouhet, V., Mougin, F., Bréchat, B. & Thiessard, F. Building a model for disease classification integration in oncology, an approach based on the national cancer institute thesaurus. en. *Journal of Biomedical Semantics* **8.** issn: 2041-1480. doi:10.1186/s13326-017-0114-4 (Dec. 2017).

194. Héja, G., Surján, G. & Varga, P. Ontological analysis of SNOMED CT. en. *BMC Medical Informatics and Decision Making* **8,** S8. issn: 1472-6947 (2008).

195. IHTSDO. *SNOMED CT Starter Guide* Feb. 2014. <http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20140222.pdf> (visited on 05/28/2016).

196. IHTSDO. *SNOMED CT Technical implementation Guide January 2015 International Release (GB English)* 2015. <http://ihtsdo.org/fileadmin/user%5C_upload/doc/download/doc%5C_TechnicalImplementationGuide%5C_Current-en-GB%5C_INT%5C_20150131.pdf?ok> (visited on 04/18/2016).

197. Bodenreider, O. *Oncology in SNOMED CT* Bethesda, Maryland, May 2015. <https://mor.nlm.nih.gov/pubs/pres/20150513-CancerBigData.pdf>.

198. Spackman, K. A., Dionne, R., Mays, E. & Weis, J. *Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED.* in *Proceedings of the AMIA Symposium* (American Medical Informatics Association, 2002), 712.

199. IHTSDO. *Mapping SNOMED CT to ICD-10 Technical Specifications* Jan. 2015.

200. *NCI Metathesaurus* 2016. <https://ncimeta.nci.nih.gov/ncimbrowser/> (visited on 04/18/2016).

201. Schuyler, P. L., Hole, W. T., Tuttle, M. S. & Sherertz, D. D. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association* **81,** 217 (1993).

202. *OWL 2 Web Ontology Language Profiles (Second Edition)* <https://www.w3.org/TR/owl2-profiles/#OWL_2_EL> (visited on 07/11/2017).

203. IHTSDO. *SNOMED CT Editorial Guide* 2016. <https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide> (visited on 04/17/2016).

204. David, J., Guillet, F. & Briand, H. *Matching directories and OWL ontologies with AROMA* in *Proceedings of the 15th ACM international conference on Information and knowledge management* (ACM, 2006), 830–831.

205. *OnAGUI - Ontology Alignment GUI download | SourceForge.net* <https://sourceforge.net/projects/onagui/> (visited on 11/02/2016).

206. Curado, M. *et al. International rules for multiple primary cancers (ICD-O Third Edition)* 2004.

207. Detwiler, L. T., Mejino, J. L. & Brinkley, J. F. From frames to OWL2: Converting the Foundational Model of Anatomy. en. *Artificial Intelligence in Medicine* **69,** 12–21. ISSN: 09333657 (May 2016).

208. Brown, S. H. *et al.* Using SNOMED CT as a reference terminology to cross map two highly pre-coordinated classification systems. eng. *Studies in Health Technology and Informatics* **129,** 636–639. ISSN: 0926-9630 (2007).

209. Bakhshi-Raiez, F., Cornet, R., Bosman, R. J., Joore, H. & de Keizer, N. F. Using SNOMED CT to identify a crossmap between two classification systems: a comparison with an expert-based and a data-driven strategy. eng. *Studies in Health Technology and Informatics* **160,** 1035–1039. ISSN: 0926-9630 (2010).

210. Schulz, S., Suntisrivaraporn, B., Baader, F. & Boeker, M. SNOMED reaching its adolescence: ontologists' and logicians' health check. en. *International Journal of Medical Informatics* **78,** S86–S94. ISSN: 13865056 (Apr. 2009).

211. Kumar, A. & Smith, B. *Oncology ontology in the NCI thesaurus* in *Conference on Artificial Intelligence in Medicine in Europe* (2005), 213–220.

212. Quix, C., Roy, P. & Kensche, D. *Automatic selection of background knowledge for ontology matching* in *Proceedings of the International Workshop on Semantic Web Information Management* (ACM, New York, USA, 2011), 5. doi:10.1145/1999299.1999304.

213. Faria, D., Pesquita, C., Santos, E., Cruz, I. F. & Couto, F. M. Automatic background knowledge selection for matching biomedical ontologies. *PloS one* **9,** e111226 (2014).

214. Alobaidi, M., Malik, K. M. & Hussain, M. Automated ontology generation framework powered by linked biomedical ontologies for disease-drug domain. *Computer Methods and Programs in Biomedicine* **165,** 117–128. ISSN: 0169-2607 (2018).

215. Coulet, A. *et al. Integration and publication of heterogeneous text-mined relationships on the Semantic Web* in *Journal of Biomedical Semantics* **2** (2011), S10.

216. Collobert, R. & Weston, J. *A unified architecture for natural language processing: Deep neural networks with multitask learning* in *Proceedings of the 25th international conference on Machine learning* (New York, USA, 2008), 160–167.

217. Kumar, A. *et al. Ask me anything: Dynamic memory networks for natural language processing* in *International conference on machine learning* (New York, USA, 2016), 1378–1387.

218. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* **13,** 55–75 (2018).

219. Castro, V. M. *et al.* Large-scale identification of patients with cerebral aneurysms using natural language processing. *Neurology* **88,** 164–168 (2017).

220. Soares, F., Villegas, M., Gonzalez-Agirre, A., Krallinger, M. & Armengol-Estapé, J. *Medical Word Embeddings for Spanish: Development and Evaluation* in *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (Association for Computational Linguistics, Minneapolis, Minnesota, USA, June 2019), 124–133. doi:10.18653/v1/W19-1916.

221. Chen, J., Jagannatha, A. N., Fodeh, S. J. & Yu, H. Ranking medical terms to support expansion of lay language resources for patient comprehension of electronic health record notes: adapted distant supervision approach. *JMIR medical informatics* **5,** e42 (2017).

222. Petrova, A. *et al.* Formalizing biomedical concepts from textual definitions. *Journal of biomedical semantics* **6,** 22 (2015).

223. Bachimont, B., Isaac, A. & Troncy, R. *Semantic commitment for designing ontologies: a proposal* in *International Conference on Knowledge Engineering and Knowledge Management* (2002), 114–121.

224. Thiao-Layel, B., Jouhet, V. & Diallo, G. *K-Ware: vers une gestion conjointe de ressources sémantiques et leurs alignements* in *6ièmes Journées Francophone sur les Ontologies* (Bordeaux, France, 2016).

225. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research,* gkw943 (2016).

226. Rappaport, N. *et al.* MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic acids research* **45,** D877–D887 (2016).

227. Dumontier, M. *et al.* The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *Journal of biomedical semantics* **5,** 14 (2014).

228. Benitez, A. A., Bourqui, R., Thebault, P. & Mougin, F. GSAn: an alternative to enrichment analysis for annotating gene sets. *bioRxiv,* 648444 (2019).

Appendix A

Original JFO article: Utilisation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie

Utilisation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie

Jean Noel NIKIEMA* — Fleur MOUGIN* —Vianney JOUHET*,**

** Equipe de Recherche en Informatique Appliquée à la Santé*
INSERM U 1219, Université de Bordeaux

{jean-noel.nikiema, fleur.mougin, vianney.jouhet }@isped.u-bordeaux2.fr

*** Service d'Information Médicale, Pôle de Santé Publique, CHU de Bordeaux*

RÉSUMÉ. *En cancérologie, la réutilisation des données est confrontée à l'hétérogénéité des terminologies. Il est ainsi nécessaire de mettre en correspondance ces dernières. L'utilisation d'une troisième terminologie comme support est une approche classique pour l'alignement de deux terminologies peu structurées. Le but de notre étude était d'utiliser la SNOMED CT comme support de connaissances afin d'aligner la CIMO3 et la CIM10, en utilisant deux approches complémentaires exploitant d'une part des mappings proposés par la SNOMED CT elle-même et d'autre part le NCI Metathésaurus. Nous avons retrouvé des mappings avec la SNOMED CT pour plus de 90% des codes CIMO3 et CIM10. Grâce à la structure de la SNOMED CT, nous avons filtré les mappings inconsistants et obtenu des mappings cohérents pour 1028/1362 codes CIMO3 et 487/852 codes CIM10. Le processus utilisant la structure de la SNOMED CT pour les alignements est tributaire de la définition logique des concepts dans la SNOMED CT. Une perspective est de prendre en compte la spécificité de ces définitions logiques afin d'améliorer le processus.*

ABSTRACT. *In oncology, the reuse of data is confronted with the heterogeneity of terminologies. It is necessary to dispose of mappings between these distinct terminologies. The semantic integration by using a third terminology as a support is a conventional approach for the alignment of two terminologies that are not very well structured. The aim of our study was to use SNOMED CT to integrate ICD-O3 and ICD10. We used two complementary resources, namely RF2 and the NCI Metathesaurus, and found mappings with SNOMED CT for over than 90% of ICD-O3 and ICD10 codes. Thanks to the structure of SNOMED CT, we filtered inconsistent mappings and obtained consistent mappings for 1028/1362 ICD-O3 codes and for 487/852 ICD10 codes. The process using the structure of SNOMED CT for establishing alignments is dependent of logical definitions of concepts in SNOMED CT. The main perspective is to take into account the specificity of these logical definitions to improve this process.*

MOTS-CLÉS : *alignement d'ontologies, interopérabilité sémantique, ontologie pivot.*

KEYWORDS: *ontology alignment, semantic interoperability, support ontology.*

1. Introduction

L’utilisation secondaire des données est un enjeu majeur en santé, notamment en cancérologie. En effet, la stratégie qui soutient la recherche et les politiques de santé pour le cancer est basée sur la surveillance des cas de cancer par les registres de cancer (Institut National du Cancer, France, 2015). Ces registres réutilisant des données provenant de différentes sources codées avec des terminologies différentes, leur intégration sémantique est un défi. La Classification Internationale des Maladies pour l’Oncologie (CIMO3) est la terminologie utilisée par les registres de cancer partout dans le monde (Fritz et al., 2008). La Classification statistique Internationale des Maladies et des problèmes de santé connexes (CIM10) est la terminologie exploitée par les sources de production de soin en France, notamment dans le cadre du Programme de Médicalisation du Système d’Information (PMSI). Au niveau international, cette dernière est également utilisée pour l’enregistrement des causes de morbidité et de mortalité (OMS, 2009). L’intégration sémantique entre la CIM10 et la CIMO3 est ainsi une nécessité pour permettre l’utilisation des données de ces différentes sources par les registres.

Cet article décrit une méthodologie exploitant la structure de la SNOMED CT pour effectuer des alignements automatiques entre la CIM10 et la CIMO3 en se basant sur un schéma général défini dans (Safar et al., 2007). Le processus d’alignement se fait ainsi en deux phases : (i) la phase d’**ancrage** qui vise à retrouver des mappings entre les codes de la CIM10 et les concepts de la SNOMED CT, ainsi qu’entre les concepts de la CIMO3 et ceux de la SNOMED CT ; (ii) la phase de **dérivation** qui consiste à retrouver les relations existant dans la SNOMED CT entre les concepts ancrés à des codes CIM10 ou CIMO3 afin de créer un pont entre les codes CIM10 et les codes CIMO3 (Figure 1).

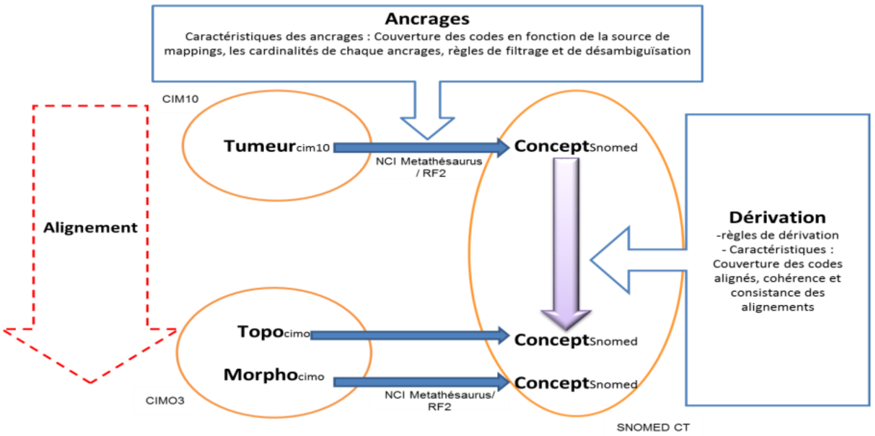

Ancrages
Caractéristiques des ancrages : Couverture des codes en fonction de la source de mappings, les cardinalités de chaque ancrages, règles de filtrage et de désambiguïsation
CIM10
Tumeurcim10
NCI Metathésaurus / RF2
ConceptSnomed
Alignement
Dérivation
-règles de dérivation
- Caractéristiques : Couverture des codes alignés, cohérence et consistance des alignements
Topocimo
ConceptSnomed
Morphocimo
ConceptSnomed
NCI Metathésaurus/ RF2
CIMO3
SNOMED CT

Figure 1. *Approche conceptuelle pour l’intégration de la CIM10 et de la CIMO3 en s’appuyant sur la SNOMED CT*

3. Matériel

3.1. *Les terminologies à aligner : la CIM10 et la CIMO3*

La CIM10 est une classification mono-axiale représentant des entités nosologiques sous la forme de codes alphanumériques (OMS, 2009). Il existe 852 codes allant de C00 à D48 décrivant les pathologies tumorales.

La CIMO3 est une classification bi-axiale décrivant d'une part les lésions histologiques des tumeurs sous forme de codes morphologiques et, d'autre part, leur(s) localisation(s) anatomique(s) sous forme de codes topographiques (Fritz et al., 2008). Les 1032 codes morphologiques comportent cinq chiffres. Les 330 codes topographiques sont des codes alphanumériques de quatre caractères.

3.2. *La terminologie pivot : la SNOMED CT*

La SNOMED CT est une terminologie multiaxiale dont tous les concepts sont organisés selon une hiérarchie sémantique à partir de 19 concepts de haut niveau (IHTSDO, 2015). Le fichier fourni par la SNOMED CT, le Release File 2 (RF2), contient les différents composants de la SNOMED CT permettant notamment d'en construire une version OWL. Le RF2 contient également des tables de correspondance avec d'autres terminologies biomédicales, en particulier avec la CIM10 et la CIMO3. La version utilisée dans notre étude est celle du 31 janvier 2016.

3.3. *Une ressource complémentaire : le NCI Metathésaurus*

Le NCI Metathésaurus (NCI Mt) est une base multi-terminologique élaborée par le National Cancer Institute américain selon le modèle du Metathésaurus de l'UMLS (Schuyler et al., 1993). Le NCI Mt contient plus de 75 terminologies biomédicales, notamment la CIM10 et la SNOMED CT. Elle intègre également des terminologies spécifiques du domaine de la cancérologie, comme la CIMO3. Le NCI Mt regroupe, selon une approche morphosyntaxique, les concepts des différentes terminologies, censés représenter la même notion sous un seul concept avec un numéro unique CUI (Concept Unique Identifier). La version du NCI Mt que nous avons utilisée est celle de juin 2013.

4. Méthodes

4.1. *La phase d'ancrage*

La phase d'ancrage a consisté en trois étapes : l'identification des mappings candidats, le filtrage des mappings incohérents et la désambiguïsation des mappings multiples.

Deux approches ont été utilisées pour la sélection des mappings candidats. Nous avons tout d'abord exploité les tables de correspondance du RF2 en ne sélectionnant que les correspondances décrites comme non obsolètes. La seconde approche, basée sur le NCI Mt, s'est attachée à identifier les CUI incluant à la fois un code SNOMED CT et un code CIM10 ou CIMO3. Après avoir calculé les couvertures des codes CIM10 et CIMO3 par la SNOMED CT, nous avons comparé les résultats obtenus par les deux approches.

Le processus de filtrage consiste à supprimer les mappings candidats inconsistants en fonction de critères auxquels devraient répondre les codes SNOMED CT impliqués dans les mappings: d'après le niveau hiérarchique, pour permettre aux concepts de décrire les mêmes notions cliniques, et le comportement tumoral. Les concepts de hauts niveau dans la SNOMED CT pour le type hierarchique sont ainsi (i) pour les codes CIM10 64572001-*Disease (disorder),* (ii) pour les codes CIMO3 morphologiques 416939005-*Proliferative mass (morphologic abnormality)* et (iii) pour les codes CIMO3 topographiques 91723000-*Anatomical structure (body structure)*. Les concepts de haut niveau pour le comportement tumoral est représenté dans le (tableau 1).

Le processus de désambiguïsation, effectué après filtrage, a pour but de choisir un seul ancrage pour les cas où un code CIM10 ou CIMO3 est mis en correspondance avec plusieurs codes SNOMED CT. Seuls les ancrages de cardinalités 1-1 (un code CIM10 ou CIMO3 mappé à un code SNOMED CT) ont été utilisés dans la phase de dérivation

Comportement tumoral		**Concept SNOMED CT de haut niveau**
CIM10	Malin primitif	372087000-*Primary malignant neoplasm (disorder)*
	Malin secondaire	128462008-*Secondary malignant neoplastic disease (disorder)* 302817000-*Malignant tumor of unknown origin or ill-defined site (disorder)*
	Hémopathies malignes	269475001-*Malignant tumor of lymphoid, hemopoietic AND/OR related tissue (disorder)*
	Tumeurs multiples	363500001-*Multiple malignancy (disorder)*
	Tumeur in situ	109355002-*Carcinoma in situ (disorder)* 127330008-*Melanoma in situ by body site (disorder)*
	Tumeur bénigne	20376005-*Benign neoplastic disease (disorder)*
	Tumeur imprévisible	118616009-*Neoplastic disease of uncertain behavior (disorder)*
CIMO3	Morphologique bénin	3898006-*Neoplasm, benign (morphologic abnormality)*
	Morphologique indéterminé	6219000-*Neoplasm, uncertain whether benign or malignant (morphologic abnormality)*
	Morphologique in situ	127569003-*In situ neoplasm (morphologic abnormality)*
	Morphologique malin primitif	86049000-*Malignant neoplasm, primary (morphologic abnormality)*
	Morphologique secondaire	14799000-*Neoplasm, metastatic (morphologic abnormality)*
	Morphologique incertain	6219000-*Neoplasm, malignant, uncertain whether primary or metastatic (morphologic abnormality)*

Tableau 1. *Concepts SNOMED CT de haut niveau associés aux codes CIM10 et aux codes CIMO3 morphologiques selon le comportement tumoral*

4.2. *La phase de dérivation*

Pour la réalisation de la dérivation, nous avons recherché tous les codes SNOMED CT équivalents ou parents de la classe anonyme correspondant à la combinaison de codes SNOMED CT ancrés à un couple de codes morphologique et topographique CIMO3 ; puis nous avons recherché les codes SNOMED CT ancrés à des codes CIM10.

5. Résultats

5.1. *Couverture des codes CIM10 et CIMO3 dans les mappings candidats*

La figure 4 présente la répartition des codes CIM10 et CIMO3 en fonction de l'approche utilisée pour établir les mappings candidats. Globalement, le chevauchement des codes CIMO3 mappé à au moins un code SNOMED CT est assez élevé. Pour la CIM10, il est plus nuancé. On retrouve par ailleurs des mappings uniquement grâce au NCI Mt pour 9% des codes CIM10, 1,5% des codes CIMO3 topographiques et 1,6% des codes CIMO3 morphologiques. Il existe également des mappings uniquement retrouvés via le RF2 pour 19% des codes CIM10, 2,4% des codes CIMO3 topographiques et 2% des codes CIMO3 morphologiques. Enfin, on note que 59 codes CIM10, 36 codes topographiques et 7 codes morphologiques ne peuvent avoir de mappings quelle que soit l'approche utilisée.

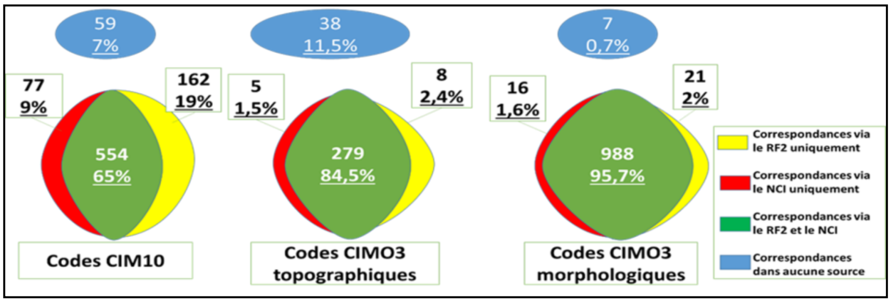


Figure 4. *Répartition des codes CIM10 et des codes CIMO3 en fonction de l'approche suivie pour établir les mappings candidats*

5.2. *Processus de filtrage et de désambiguïsation de la phase d'ancrage*

Le tableau 2 représente l'impact du processus de filtrage et de désambiguïsation à chaque étape d'après le nombre de codes CIM10 et CIMO3 impliqués dans les mappings de cardinalités 1-0, 1-1 et 1-N.

Etapes		Cardinalités	CIMO3				CIM10	
			Topographique		Morphologique			
			RF2	NCI	RF2	NCI	RF2	NCI
Identification des mappings candidats		1-0	43	46	23	28	136	221
		1-1	4	132	960	539	79	516
		1-N	283	152	49	465	637	115
Filtrage selon le type hiérarchique		1-0	1	29	1	7	0	11
		1-1	4	125	959	847	79	572
		1-N	282	130	49	150	637	48
Filtrage selon le comportement tumoral		1-0			48	56	50	305
		1-1			912	838	159	288
		1-N			48	103	507	27
Désambiguïsation	Supprimés		44	75	72	91	186	537
	Validés	1-1	131	184	957	879	448	302
	A traiter	1-N	155	71	3	62	218	13

Tableau 2. *Nombre de codes CIM10 et CIMO3 impliqués dans des mappings à chaque étape du processus de filtrage et de désambiguïsation.*

5.3. *Phase de dérivation*

On retrouve que 203/437 codes CIM10 désambiguïsés peuvent être dérivés avec 127/127 codes CIMO3 topographiques désambiguïsés et 888/901 codes CIMO3 morphologiques désambiguïsés. Un exemple d'alignement de cardinalités 1-1 est le code CIM10 d'hémopathie maligne C91-*Leucémie lymphoïde* avec la combinaison de codes CIMO3 C42.1-*Moelle osseuse* et 9820/3-*Leucémie lymphoïde, SAI.*

6. Discussion

La CIM10 et la CIMO3 n'ont pas la même structure et ne représente pas les même notions clinique d'où l'intérêt de rechercher des liens sémantiques entres les codes CIM10 et CIMO3 par une approche utilisant une troisième ressource termino-ontologique de support plutôt que de procéder à des mappings direct entre ces codes. Nous avons utilisé deux approches existantes (RF2 et NCI Mt) pour la mise en évidence de nos ancrages, cependant les autres méthodes, tels que AROMA (David et al., 2006), et applications de mappings tels que ServOMap (Diallo et al., 2014) peuvent être utilisées pour la création des ancrages. Le choix de la SNOMED CT comme ressource termino-ontologique de support semble être judicieux selon notre étude puisque nous avons mis en évidence une couverture globalement élevée des codes CIM10 et CIMO3 (Nikiema et al., 2016). Le NCI thésaurus a été utilisé comme alternative à la SNOMED CT pour l'alignement de la CIM10 et de la CIMO3 dans (Jouhet et al., 2014). Une perspective serait de comparer la qualité des alignements obtenus dans notre étude avec ceux de (Jouhet et al., 2014).

Le processus de filtrage et de désambiguïsation des ancres nous ont permis d'améliorer la qualité des ancres trouvés mais également de gérer les contradictions entre les différentes approches (RF2 et NCI Mt). Cependant ces processus ainsi que la phase dérivation est tributaire de la qualité des définitions logique des concepts de la SNOMED CT. On retrouve alors le code CIMO3 morphologique malin primitif 8151/3-*insulinome malin* se retrouve aligner à des codes CIM10 de tumeur bénigne car il est ancré à 20955008-*Insulinome malin (morphologic abnormality*) qui est décrit dans la SNOMED CT comme descendant 3898006-*tumeur bénigne (morphologic abnormality).* Cet exemple met en évidence une utilisation incontrôlée de la relation is_a dans la SNOMED CT, appelée **is_a overloading** (Bondenreider et al., 2007). Nous avons ainsi décelé, grâce à nos choix méthodologiques, des pistes à explorer pour l'intégration sémantique de tous les codes CIMO3 et CIM10 via la SNOMED CT en prenant en compte les spécificités de ces trois ressources termino-ontologiques. De plus, ce travail nous a menés indirectement à analyser la qualité de la structure de la SNOMED CT et ainsi à participer à l'audit de cette ressource termino-ontologique.

Bibliographie

Bodenreider O, Smith B, Kumar A, Burgun A. « Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies». *Artif Intell Med.*,2007;p183-95

David, J., Guillet, F., and Briand, H.. « Matching directories and OWL ontologies with AROMA». In *International Conference on Information and Knowledge Management, (ACM)*,2006, p830–831

Fritz A., et OMS. *Classification internationale des maladies pour l'oncologie*. Genève: Organisation mondiale de la santé, 2008.

IHTSDO. « SNOMED CT® Technical implementation Guide January 2015 International Release (GB English) », 2015.

Institut National du Cancer, France. « Plan cancer 2014-2019 Guérir et prévenir les cancers: donnons les mêmes chances à tous, partout en France », 2015.

Jouhet, V., B. Bréchat, F. Mougin, et F. Thiessard. « Intégration de terminologies diagnostiques en cancérologie. » Revue d'Épidémiologie et de Santé Publique 62, n°S5, septembre 2014, p. 185

Nikiema, J.N., Jouhet V., et Mougin F. « Evaluation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie ». In *Atelier Intelligence Artificielle et Santé (IC2016)*. Montpellier, 2016.

OMS. *Classification statistique internationale des maladies et des problèmes de santé connexes, Dixième version*, 2009.

Safar B., Reynaud C., et Calvier F. « Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire ».In *1ères Journées Francophones sur les Ontologies (JFO 2007)*, 2007, p 21–35.

Schuyler, Peri L., William T. Hole, Mark S. Tuttle, et David D. Sherertz. « The UMLS Metathesaurus: representing different views of biomedical concepts. » *Bulletin of the Medical Library Association 81, n° 2*, 1993, p. 217-222.

Diallo G. « An effective method of large scale ontology matching ». *J Biomed Semant,* 2014, p44.

Appendix B

Original JBI article: Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts

Journal of Biomedical Informatics 74 (2017) 46–58

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts

Jean Noël Nikiema [a,*], Vianney Jouhet [a,b,1], Fleur Mougin [a,1]

[a] *Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, Team ERIAS, UMR 1219, Bordeaux F-33000, France*
[b] *CHU de Bordeaux, Pole de sante publique, Service d'information medicale, Unit IAM, F-33000 Bordeaux, France*

ARTICLE INFO

Article history:
Received 12 May 2017
Revised 21 July 2017
Accepted 23 August 2017
Available online 24 August 2017

Keywords:
Semantic integration
Biomedical terminologies
Logical definitions
Post-coordination
SNOMED CT
ICD-10
ICD-O3

ABSTRACT

In oncology, the reuse of data is confronted with the heterogeneity of terminologies. It is necessary to semantically integrate these distinct terminologies. The semantic integration by using a third terminology as a support is a conventional approach for the integration of two terminologies that are not very structured. The aim of our study was to use SNOMED CT for integrating ICD-10 and ICD-O3. We used two complementary resources, mapping tables provided by SNOMED CT and the NCI Metathesaurus, in order to find mappings between ICD-10 or ICD-O3 concepts and SNOMED CT concepts. We used the SNOMED CT structure to filter inconsistent mappings, as well as to disambiguate multiple mappings. Based on the remaining mappings, we used semantic relations from SNOMED CT to establish links between ICD-10 and ICD-O3. Overall, the coverage of ICD-O3 and ICD10 codes was over 88%. Finally, we obtained an integration of 24% (203/852) of ICD-10 concepts with 86% (888/1032) of ICD-O3 morphology concepts combined to 39% (127/330) of ICD-O3 topography concepts. Comparing our results with the 23,684 ICD-O3 pairs mapped to ICD-10 concepts in the SEER conversion file, we found 17,447 pairs of ICD-O3 concepts in common among which 11,932 pairs were integrated with the same ICD-10 concept as the SEER conversion file. The automated process leverages logical definitions of SNOMED CT concepts. While the low quality of some of these definitions impacted negatively the integration process, the identification of such situations made it possible to indirectly audit the structure of SNOMED CT.

1. Introduction

Secondary use of biomedical data is a major issue because, nowadays, it supports health systems' improvement and a better understanding of diseases and treatments. Indeed, patient information collected for care delivery can also be used for research and billing purposes as well as certification and accreditation of health facilities, evidence-based medicine and business applications [1]. Secondary use hence opens perspectives for applying data mining approaches to the biomedical domain. These approaches are promising to "*greatly expand the capacity to generate new knowledge*" and "*help translate personalized medicine initiatives into clinical practice by offering the opportunity to use analytical capabilities that can integrate systems biology (e.g., genomics) with EHR data*" [2]. Through the identification of cancer cases by registries, oncology is an area where the secondary use of health data is particularly important [3]. Indeed, the goal of cancer registries is to track all cases of cancer occurring in a defined population. The cancer data are continuously and systematically collected from various healthcare facilities (hospital, pathology laboratory, etc.). The collected data consist of sociodemographic information about the patients, as well as clinical and histopathological characteristics of the cancer that is being studied. Cancer registries thus allow follow-up of patients diagnosed with cancer and provide statistical results on the outcomes of the corresponding disease (mortality, results of therapy, etc.) [4]. The collected data are also used for epidemiological research on cancer incidence and determinants, as well as for supporting evidence for health policies on diagnosis, prevention and cancer treatment [5]. Therefore, cancer monitoring requires the concomitant use of data coming from different sources with the difficulty that these data are potentially encoded according to different terminological resources [6]. Indeed, for encoding data, cancer registries use the third edition of the International Classification of Diseases for Oncology (ICD-O3) [7], while, in France for instance, the medical data in hospitals are encoded (for billing purpose through the PMSI "Programme de Médicalisation du Systéme d'Information" [8]) according to the tenth revision of the International statistical Classification of Diseases and related

[*] Corresponding author.
E-mail address: jean.nikiema@u-bordeaux.fr (J.N. Nikiema).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.jbi.2017.08.013
1532-0464/© 2017 Elsevier Inc. All rights reserved.

health problems (ICD-10). In addition, morbidity and mortality causes are internationally recorded using ICD-10 [9]. The achievement of the objectives of cancer registries thus requires, among other things, the joint use of ICD-10 and ICD-O3, making their semantic integration essential in France and all around the world.

ICD-10 and ICD-O3 are knowledge resources called terminologies. A *terminology* is the product of a science that aims to make an inventory of a given domain's concepts and the terms that designate them [10]. On the other hand, an *ontology* is defined as "*a formal and explicit specification of a shared conceptualization*" according to Studer et al. [11]. In the biomedical domain, the notions of terminology and ontology are frequently confused. To provide a general framework applicable to alignment methods whatever the level of structuring of knowledge resources to be aligned, we will use the term of termino-ontological resource (TOR). As described by Jouhet et al., ICD-10 and ICD-O3 exhibit structural heterogeneities. As a result, it is not possible to find equivalences between concepts of these two terminologies. In this context, linking ICD-10 and ICD-O3 requires a true reconciliation of the concepts they describe. The notion of *reconciliation* in our study consists in identifying any type of relation that can exist between two concepts (equivalent and subsumption relations and, in case of disjunction, the appropriate non-hierarchical relation). In the remaining part of this paper, we will use the term *integration* in order to denote this reconciliation process between all the concepts coming from different TORs.

In Section 2, existing approaches for aligning distinct TORs are presented. Then, a description of the characteristics of ICD-10 and ICD-O3 is provided before the justification of our methodological choices for performing their semantic integration. Lastly, our analytical framework is featured. In Sections 3 and 4, we present the materials that we used and the methods that we developed for the semantic integration of ICD-10 and ICD-O3. Finally, we present and discuss our results in Sections 5 and 6.

2. Background

2.1. Related work

In the literature, to carry out an alignment between two TORs, authors generally proceed to the identification of mappings. Mappings are formal expressions of correspondences (equivalence, subsumption and disjunction) between two concepts [12]. Several methods have been proposed for establishing mappings in order to perform the alignment of TORs. Saitwal et al. grouped these techniques into four methodological approaches: (i) manual approaches, (ii) morphosyntactic approaches, (iii) approaches based on semantic features of concepts (subsumption, roles, etc.) and the structure of TORs to be aligned and, finally, (iv) approaches using a third TOR as background knowledge [13].

The various methodological approaches mentioned above have been used individually or combined and their utilization often led to the development of softwares for implementing these approaches. Some electronic tools have been developed to serve as support for the manual creation of mappings between TOR concepts. Examples of manual approaches followed in the biomedical domain include the work of Giannangelo et al., who created mappings between concepts of ICD-10 and SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms) [14], and the work of Souvignet et al., who established mappings between concepts and relations of PS-CAST (Patient Safety Categorical Structure), an ontology made by the WHO (World Health Organization), and BFO (Basic Formal Ontology), an upper level ontology [15]. Mappings were validated by multiple experts in a consensual way in [14] while the logical consistency of mappings was checked using the OWL DL reasoner Hermit in [15].

Beyond the biomedical domain, the international Ontology Alignment Evaluation Initiative (OAEI) evaluates and compares systems and algorithms that combine several automated mapping techniques. OAEI is a benchmarking initiative started in 2004 with the participation of four systems. During the 2015 campaign, 22 systems participated [16]. As Shvaiko et al. have shown by an analysis of the algorithms of systems participating in the OAEI campaigns, most systems begin the creation of mappings with morphosyntactic approaches, *i.e.*, by searching for string and linguistic similarities between concept labels [17]. This is the case in SAMBO [18], QODI [19], AgreementMaker [20] and ServOMap [21].

SAMBO, ServOMap and QODI also use the structure of the two input TORs for finding additional mappings, based on initial mappings which have been discovered at the lexical level. In SAMBO, mappings are created between two concepts which lie in a similar position with respect to *is_a* (*i.e.*, subsumption) or *part_of* relationships according to the mappings identified morphosyntactically. ServOMap establishes similarity between concepts having the same structural proximity (*i.e.*, parents, siblings and descendants) according to already mapped concepts. QODI calculates similarity measures according to paths between concepts. Authors select a specific path between two concepts in a first ontology. Then, they compare this path with different possible paths present in the second ontology. The comparison is based on the similarity between: (i) source concepts of each path, (ii) datatype properties, (iii) labels of the concepts located between the source concept and the last concept within each path, and on a penalty for path length difference. It is important to notice that the use of the structure to find mappings between two TORs requires that the latter have a fairly high level structure.

Some mapping tools use additional knowledge resources for the creation of mappings according to similarity measures between concepts in this resource. AgreementMaker [20] uses Wordnet [22], a lexical database for English, and UBERON [23], a multi-species anatomy ontology. SAMBO uses also two external resources. The first resource used by SAMBO is the UMLS (Unified Medical Language System) Metathesaurus®, a multi-terminological system containing more than 200 biomedical terminologies, for finding similarities between concepts of TORs which are included in this resource. The second resource used by SAMBO is PubMed for calculating similarity measures based on the co-occurrence of TOR concepts to be mapped in a set of PubMed abstracts.

In addition with the four approaches described in [13], another approach for aligning TORs is to leverage the post-coordination principle. Post-coordination is the combination of concepts to express a notion which is not already described in a TOR [24]. Within a given TOR or in different TORs, a notion can be represented thanks to both pre-coordinated concepts (*e.g., Breast cancer*) and post-coordinated expressions (*e.g., Breast + Malignant neoplasm*). For creating a bridge between pre-coordinated concepts and post-coordinated expressions within a unique TOR, the strategy proposed by Dolin et al. consists in providing a canonical form to pre-coordinated concepts and post-coordinated expressions through Health Level 7 Reference Information Model (HL7 RIM) [25]. For the creation of mappings between pre-coordinated concepts and post-coordinated expressions existing in different TORs, Dhombres et al. proposed a method that was carried out to increase the coverage of HPO (Human Phenotype Ontology) concepts mapped to SNOMED CT concepts [26]. Authors developed an algorithm for identifying each term that represents a clinical notion within HPO concept labels (*e.g.*, the HPO label "*abnormality of the lip*" contains "*abnormality*" as disorder and "*lip*" as *anatomical structure*). These terms were then mapped to SNOMED CT concepts according to a morphosyntactic method. The resulting SNOMED CT concepts were combined (through post-coordination) to represent some disorders that correspond to the HPO concept (*e.g.*, still for the HPO concept "*abnormality of the lip*", the corresponding post-coordination expression proposed by authors is the following: 64572001-*Disease (disorder)* + 363698007-*Finding site (attribute)* + 48477009-*Lip structure (body structure)*). This method was thus able to find mappings between pre-coordinated concepts from HPO with post-coordinated expressions in SNOMED CT. Note that this method requires the concept label to be interpretable with a sophisticated syntax and a structure allowing the automatic post-coordination of TORs to be aligned.

Thus, in the literature, approaches which aim at creating automatically mappings, whatever the strategy used, are mainly concerned with finding similarity measures between concepts coming from two different TORs and mostly rely on their labels. These approaches try to identify equivalences or subsumption relations between concepts, which corresponds to alignment [12]. However, it is sometimes not possible to find such mappings when TORs represent concepts from disjoint semantic categories. In this case, because they cannot be aligned (no direct mappings exist between concepts), TORs need to be integrated by reconciling their concepts [27,28]. In this frame, our goal is to provide an integration process of ICD-10 and ICD-O3 for a true reconciliation of their concepts, which depends neither on a great structuration of the TORs to be integrated, nor on the expressivity of their concept labels.

2.2. Characteristics of ICD-10 and ICD-O3

2.2.1. ICD-10

ICD-10 is a classification maintained by the WHO for representing nosologic entities through alphanumeric codes. The nosologic entities are autonomous in their determinism. They are also consistent in their clinical manifestations and organized according to their similarities and contrasts. Consequently, ICD-10 concepts are disjoint. Chapter II of ICD-10 is dedicated to tumors, in which 852 alphanumeric codes range from C00 to D48:

- C00-C97: concepts of malignant neoplasms.
  - C00-C75: concepts of malignant neoplasms, stated or presumed to be primary, of specified sites, except for lymphoid, haematopoietic and related tissues.
  - C76-C80: concepts of malignant neoplasms of ill-defined, secondary and unspecified sites.
  - C81-C96: concepts of malignant neoplasms, stated or presumed to be primary, of lymphoid, haematopoietic and related tissues.
  - C97: concept of malignant neoplasms of independent (primary) multiple sites.
- D00-D09: concepts of in situ neoplasms.
- D10-D36: concepts of benign neoplasms.
- D37-D48: concepts of neoplasms having an uncertain or unknown behavior.

The classification of tumors is mainly made by site, and in very large groups, depending on the behavior of the tumor.

2.2.2. ICD-O3

ICD-O3 is a biaxial classification describing, on the one hand, histological lesions of tumors concepts (morphology), and on the other hand, their anatomical location(s) concepts (topography). The 1032 morphology codes start with the letter "M-" followed five digits between M-8000/0 and M-9989/3. The first four digits represent the specific histologic term. The fifth digit, behind the slash (/), indicates the behavior of the tumor, *i.e.* whether it is primary malignant (/3), secondary malignant (/6), benign (/0), in situ (/2), with an uncertain or unknown behavior (/9) or undetermined behavior (/1). The 330 topography codes are composed of four characters and range from C00.0 to C80.9.

2.3. Methodological choice for the semantic integration of ICD-10 and ICD-O3

ICD-10 and ICD-O3 are two classifications that differ by:

- The clinical concepts that they describe: ICD-10 represents diseases whereas ICD-O3 describes histological lesions and anatomical sites.
- Their structure: ICD-10 is mono-axial while ICD-O3 is biaxial.

Each ICD-10 concept is used independently to record health data and expresses a diagnosis as a whole, thus corresponding to a pre-coordinated concept. In contrast, an ICD-O3 morphology concept must be associated to an ICD-O3 topography concept in order to express the complete diagnosis to be recorded. There are no combination rules in ICD-O3. Thus, all combinations of ICD-03 topography and morphology concepts are potentially allowed. ICD-O3 concepts thus need to be combined for finding mappings with ICD-10 concepts. The link between ICD-10 and ICD-O3 can be made by describing a cancer disease (coded in ICD-10) in terms of its manifestation (ICD-03 morphology concept) and its localization (ICD-O3 topography concept). The semantic integration of ICD-10 and ICD-O3 thus requires the creation of a link between ICD-10 pre-coordinated concepts and post-coordinated expressions corresponding to a combination of ICD-O3 morphology and topography concepts.

Because ICD-10 and ICD-O3 are large, their manual reconciliation would be a long and tedious task [29]. On the other hand, morphosyntactic approaches exploiting the concept labels do not take into account the "pre-coordinated" and "post-coordinated" characteristics of ICD-10 and ICD-O3. Finally ICD-10 and ICD-O3 are not described in a formal language. They are not ontologies but just classifications describing disjoint concepts. Thus, integrating them on the basis of their semantic features cannot take into account post-coordination issues. As a result, based on the classification proposed by Saitwal et al. [13], we propose to use background knowledge available in a support TOR in order to create a method for integrating ICD-O3 and ICD-10. The method using a TOR as a support resource for TORs' integration is particularly relevant when they are weakly structured or limited to simple classification hierarchies [30]. Indeed, this method allows to compensate this weakness by articulating concepts of TORs to be integrated according to the support TOR, making it possible to automatically find correspondences or to analyze their quality.

2.4. Analytical framework

In order to integrate ICD-10 and ICD-O3 by using a support TOR, we have defined a conceptual framework based on the general patterns described by Safar et al. [30]. It consists in two stages (Fig. 1): (1) the anchoring stage, which aims to generate candidate mappings (called anchors) between concepts of the TORs to be integrated and concepts of the support TOR, (2) the derivation stage, which consists of identifying links between the concepts participating in the anchors within the support TOR so that concepts from the TORs to be integrated can be related to each other.

To implement this framework, it was first necessary to select an appropriate support TOR. The latter must be able to describe the appropriate relations between the notions of tumor diseases, histological lesions and anatomical localizations. Its structure must also allow logical inference because ICD-10 and ICD-O3 are large. Indeed, the automatic deduction of relations and constraints

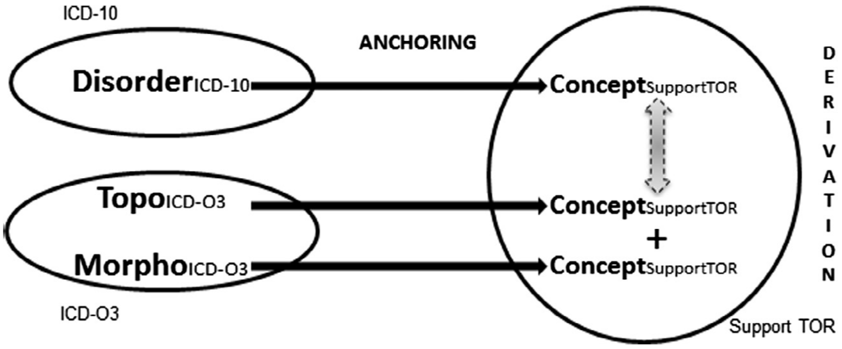


Fig. 1. Analytical framework: description of general patterns used to integrate ICD-10 and ICD-O3 using a support TOR. Two stages for this semantic integration: (1) the anchoring stage generates candidate mappings between ICD-10 and ICD-O3 concepts and the support TOR, (2) the derivation stage identifies links between the concepts of the support TOR which participate in the anchors.

existing between the concepts of the support TOR has to be possible, without these relations and constraints being specifically expressed by TOR creators [31]. The external resource, which is the most commonly used support TOR within the biomedical domain, is the UMLS Metathesaurus. For instance, it was exploited to align GALEN (Generalised Architecture for Languages, Encyclopaedias and Nomenclatures in medicine) and TAMBIS (Transparent Access to Bioinformatics Information Sources) [32], MedDRA and SNOMED CT [33], as well as CCC (Clinical Care Classification) and NANDA-I (North American Nursing Diagnosis Association-International) [34]. Another biomedical TOR, which has been used as a support for aligning the ATC (Anatomical Therapeutic Chemical) and the MeSH (Medical Subject Headings), is RxNorm [35]. In the specific domain of oncology, the external resource which is often used is the NCI Metathesaurus [36,37], while Jouhet et al. have exploited the NCI thesaurus [38].

3. Materials

3.1. Support TOR: SNOMED CT

For our study, we chose SNOMED CT for integrating ICD-10 and ICD-O3 (Fig. 1). SNOMED CT is one of the most descriptive biomedical knowledge resources exhibiting ontological characteristics [39]. SNOMED CT is based on three types of components: (i) *concepts* which represent a clinical meaning and have a unique identifier (SCTID), (ii) *descriptions* which represent labels of these concepts and (iii) *relations* which are binary links between concepts [40]. SNOMED CT is described in description logics (DL) [41] with over 300,000 concepts organized according to a hierarchy rooted by 19 high-level classes, among which *Clinical finding* (which has *Disease* among its descendants) and *Body structure* (which has *Proliferative mass* and *Anatomical structure* among its descendant concepts) are of special interest for this work. Each SNOMED CT concept has at least one subsumption relation with another SNOMED CT concept.

SNOMED CT also associates logical definitions to most of its concepts. This logical definition of is composed of other SNOMED CT concepts and relations [42]. In SNOMED CT, it is thus possible to describe a tumor thanks to the semantic link *Associated morphology* relating to a concept describing its histologic lesion as well as the semantic link *Finding site* relating to a concept describing its anatomical location [43]. Since 2002, SNOMED CT has introduced a particular relation that accompanies other relations, called *role_group*. This relation has been introduced for better describing diseases which have several sites or morphological abnormalities. More precisely, the *role_group* relation enables to describe the morphological lesion which is associated with each anatomical site [44,45]. For example, according to the Manchester syntax [46], the logical definition of the concept *Tetralogy of Fallot* (which is a cardiac malformation characterized by different anomalies, which affect multiple anatomical sites) is:

Ventricular septal defect **AND** Right ventricular hypertrophy
  AND Pulmonic valve stenosis **AND**
Overriding aorta **AND** Congenital abnormality of ventricles
  and ventricular septum
  AND role_group SOME (Associated morphology SOME
  Congenital failure of fusion **AND**
    Finding site SOME Interventricular septum structure **AND**
    Occurrence SOME Congenital)
  AND role_group SOME (Associated morphology SOME
  Stenosis **AND**
    Finding site SOME Pulmonary valve structure)
  AND role_group SOME (Associated morphology SOME
  Overriding structures **AND**
    Finding site SOME Thoracic aorta structure)
  AND role_group SOME (Associated morphology SOME
  Hypertrophy **AND**
    Finding site SOME Entire right ventricle)
  AND role_group SOME (Associated morphology SOME
  Developmental anomaly **AND**
    Finding site SOME Aortic structure **AND**
    Occurrence SOME Congenital)

3.2. Mapping resources for integrating ICD-10 and ICD-O3

The SNOMED CT mapping tables (SNCTmt). SNOMED CT provides a file which contains, among others, mapping tables between SNOMED CT concepts and ICD-10 as well as ICD-O3 concepts [47]. These mappings have been established manually and their purpose was to find, for a given SNOMED CT code, the corresponding ICD-10 code(s) or ICD-O3 code(s).

The NCI Metathesaurus (NCI Mt). The NCI Mt is a multi-terminology database integrating around 100 biomedical TORs related to cancer [48]. ICD-10, ICD-O3 and SNOMED CT used in this study are included within the NCI Mt. Like in the UMLS Metathesaurus, each concept in the NCI Mt has a unique identifier, named Concept Unique Identifier (CUI), which clusters the codes from distinct TORs supposed to represent the same notion. This clustering has been performed according to a morphosyntactic approach [49].

4. Methods

For the integration process, a preliminary stage has been performed to recover the exhaustive list of ICD-10 codes “from C00 to D48”, as well as the ICD-O3 codes from the NCI Mt. This list has been rid of ICD-10 and ICD-O3 header codes (*e.g.*, C00-C97 Malignant neoplasms) because, in practice, they are not used for diagnostic coding.

Then, we exploited the structure of SNOMED CT to which we applied the ELK reasoner [50]. A reasoner is an algorithm that can infer logical consequences from explicit assertions. We chose to use ELK because it has been described efficient for performing a quick and efficient ranking of SNOMED CT [50,51]. Thus, we used the ELK reasoner trough the OWLAPI 3.5.0 at the anchoring and derivation stages.

4.1. Anchoring stage

The anchoring stage consists of three steps: identifying candidate mappings for anchoring, filtering anchors and disambiguating multiple anchors (Fig. 2). For the two last steps, we used ELK to infer the whole SNOMED CT structure so that subsumption relations which are not explicitly stated between some SNOMED CT concepts are also available.

4.1.1. Identifying candidate mappings

Two resources were used for the selection of candidate mappings. We first used the SNCTmt by selecting only the anchors described as not obsolete. The second mapping resource, namely the NCI Mt, was exploited to identify the CUI including both a SNOMED CT code (SCTID) and an ICD-10 or ICD-O3 code (Fig. 3).

4.1.2. Filtering anchors

To eliminate inconsistent anchors, two sub-steps were performed: (i) a filtering according to the SNOMED CT hierarchy and (ii) a filtering according to the tumor behavior.

The filtering according to the hierarchy aimed to remove the anchors which involved concepts that do not represent the same general clinical notions. Thus, the anchors were considered as inconsistent in the following cases:

- for ICD-10 concepts (which represent diseases): if the mapped SNOMED CT concept was not a descendant of the concept 64572001-*Disease (disorder)*,
- for ICD-O3 morphology concepts (which represent histologic lesions): if the mapped SNOMED CT concept was not a descendant of the concept 416939005-*Proliferative mass (morphologic abnormality)*,
- for ICD-O3 topography concepts (which represent anatomical localizations): if the mapped SNOMED CT concept was not a descendant of the concept 91723000-*Anatomical structure (body structure)*.

The filtering according to the tumor behavior was applied only to anchors in which ICD-10 concepts and ICD-O3 morphology concepts participated. This step consisted in the reconciliation of the different classes of tumor behavior found in the structure of the ICD-10 or within the ICD-03 morphology axis with those represented in SNOMED CT. In practice, all anchors involving concepts that do not describe the same kind of tumor behavior were removed. For that, we have identified the high-level SNOMED CT concepts that correspond to classes of tumor behavior which are represented within the ICD-10 structure and within the morphology axis of ICD-O3. The list of high-level SNOMED CT concepts chosen for each class is presented in Table 1. Some classes of tumor behavior in ICD-10 have multiple corresponding SNOMED CT concepts because these classes represent distinct notions that are not grouped together within SNOMED CT (*e.g.*, the high-level SNOMED CT concepts chosen for the ICD-10 class “Tumor in situ” are 109355002-*Carcinoma in situ (disorder)* and 127330008-*Melanoma in situ by body site (disorder))*.

4.1.3. Disambiguating multiple anchors

The objective of the disambiguation process is to propose the best anchor(s) when several SNOMED CT concepts are mapped to a single ICD-10 or ICD-O3 concept. For that, we examined the

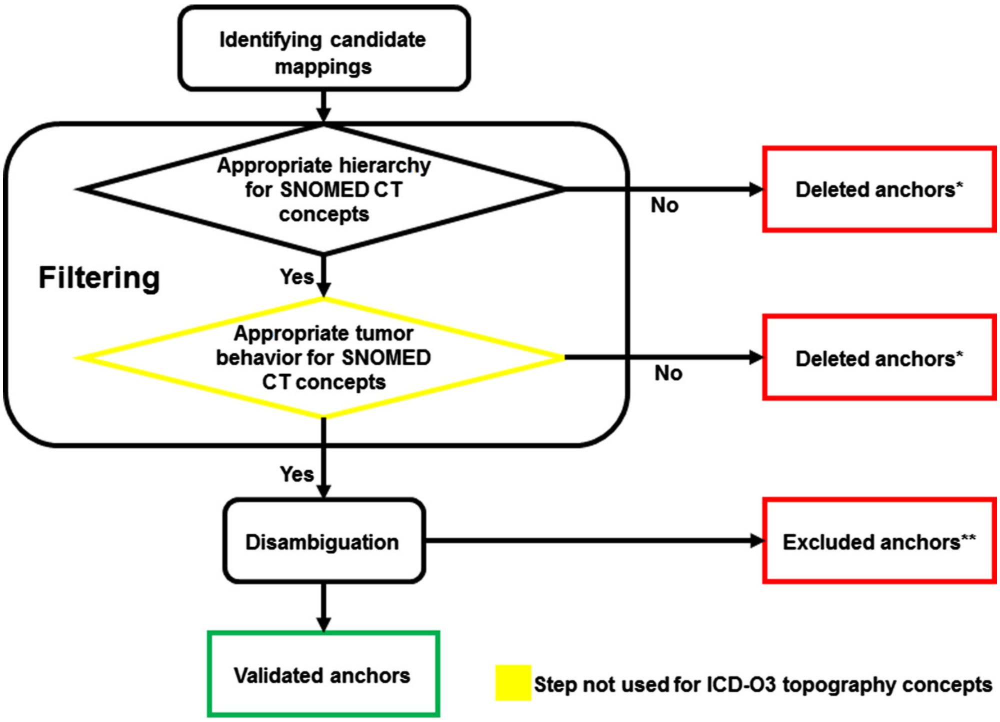

Identifying candidate mappings
Appropriate hierarchy for SNOMED CT concepts
No
Deleted anchors*
Filtering
Yes
Appropriate tumor behavior for SNOMED CT concepts
No
Deleted anchors*
Yes
Disambiguation
Excluded anchors**
Validated anchors
Step not used for ICD-O3 topography concepts

Fig. 2. The three steps of the anchoring stage: (1) the identification of candidate mappings, (2) the filtering step, which consists in deleting anchors involving concepts that do not describe the same clinical notions, (3) the disambiguation step for excluding anchors involving a unique ICD-10/ICD-O3 concept and multiple SNOMED CT concepts. ⁎Deleted anchors are erroneous mappings. ⁎⁎Excluded anchors are correct mappings which do not denote equivalences.

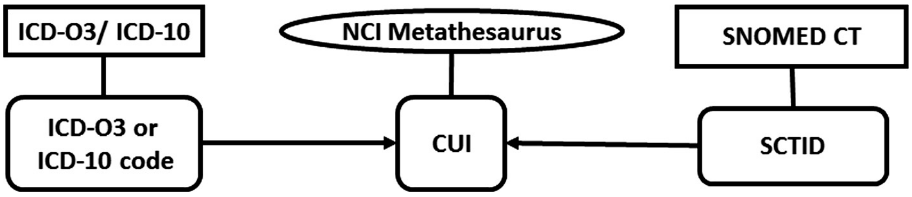


Fig. 3. Identifying candidate mappings within the NCI Metathesaurus.

Table 1
High-level SNOMED CT concepts corresponding to classes of tumor behaviors in ICD-10 and ICD-O3.

Classes of tumor behavior		Corresponding high-level SNOMED CT concept(s)
ICD-10	Primary malignant (C00-C75)	372087000-*Primary malignant neoplasm (disorder)*
	Secondary malignant (C76-C80)	128462008-*Secondary malignant neoplastic disease (disorder)* 302817000-*Malignant tumor of unknown origin or ill-defined site (disorder)*
	Haematological malignancy (C81-C96)	269475001-*Malignant tumor of lymphoid, hemopoietic AND/OR related tissue (disorder)*
	Multiple tumors (C97)	363500001-*Multiple malignancy (disorder)*
	Tumor in situ (D00-D09)	109355002-*Carcinoma in situ (disorder)* 127330008-*Melanoma in situ by body site (disorder)*
	Benign tumor (D10-D36)	20376005-*Benign neoplastic disease (disorder)*
	Unpredictable tumor (D37-D48)	118616009-*Neoplastic disease of uncertain behavior (disorder)*
ICD-O3	Benign (/0)	3898006-*Neoplasm, benign (morphologic abnormality)*
	Undetermined behavior (/1)	86251006-*Neoplasm, uncertain whether benign or malignant (morphologic abnormality)*
	Uncertain or unknown tumor behavior (/9)	6219000-*Neoplasm, malignant, uncertain whether primary or metastatic (morphologic abnormality)*
	In situ morphology (/2)	127569003-*In situ neoplasm (morphologic abnormality)*
	Primary malignant morphology (/3)	86049000-*Malignant neoplasm, primary (morphologic abnormality)*
	Secondary malignant morphology (/6)	14799000-*Neoplasm, metastatic (morphologic abnormality)*

existence of subsumption relations between these SNOMED CT concepts. We then kept only the anchor(s) involving the SNOMED CT concept(s) being the most generic (*i.e.*, situated at the highest level in the hierarchy).

Disambiguation does not reduce the number of ICD-10 or ICD-O3 concepts involved in anchors but only the number of SNOMED CT concepts mapped to them. The same disambiguation process was applied to the anchors of each resource (first step), and to the pooling of anchors obtained at the first step (second step).

More precisely, we first addressed the disambiguation of the anchors coming from the SNCTmt independently from those coming from the NCI Mt. This step was intended to harmonize anchors within each of these two resources. At the second step, the disambiguated anchors obtained from the two resources were pooled and a second disambiguation was performed, when needed. Indeed, pooling anchors lead to two situations. Given an ICD-10/ ICD-O3 concept:

- Anchor(s) retrieved by the two resources was (were) the same or only one resource retrieved the anchor(s). In this situation, no additional disambiguation was needed.
- Anchors retrieved by the two resources were different (distinct SNOMED-CT concepts). In this situation, the disambiguation process was performed over pooled anchors. Thus, if one of the SNOMED CT concepts involved in the multiple anchors was more general than others, this step allowed to transform an 1–N anchor into an 1–1 anchor.

4.1.4. Evaluation of steps of the anchoring stage

In order to evaluate the methods used during the anchoring stage, we first estimated the coverage of ICD-10 and ICD-O3 concepts within anchors and compared the results obtained through the SNCTmt and the NCI Mt. Then, to assess the impact of each step of the anchoring stage, we calculated the number of anchors obtained for each ICD-10 and ICD-O3 concept and having the following cardinalities before and after each step:

- 1–1 anchors: an ICD-10 or ICD-O3 concept mapped to a single SNOMED CT concept.
- 1–N anchors: an ICD-10 or ICD-O3 concept mapped to more than one SNOMED CT concept.
- 1–0 anchors: an ICD-10 or ICD-O3 concept which could not be mapped to any SNOMED CT concept.

4.2. Derivation stage

4.2.1. Derivation method

This step consisted in identifying the relations existing between SNOMED CT concepts participating in the anchors in order to deduce correspondences between ICD-10 concepts and combinations of an ICD-O3 morphology concept and an ICD-O3 topography concept (Fig. 4). Only 1–1 anchors obtained at the end of the anchoring stage were used for the derivation stage. Therefore, each possible pair of anchored ICD-O3 morphology and topography concepts corresponds to a unique pair of SNOMED CT concepts. For each of these SNOMED CT concept pairs, we looked for the SNOMED CT concepts of disease (equivalent concept or, failing that, parent concepts) that have the appropriate semantic link with each element of the pair (*i.e.*, a *finding_site* relationship with anatomical structures and an *associated_morphology* relationship with histological lesions). Toward this end, we automatically generated DL-queries which were

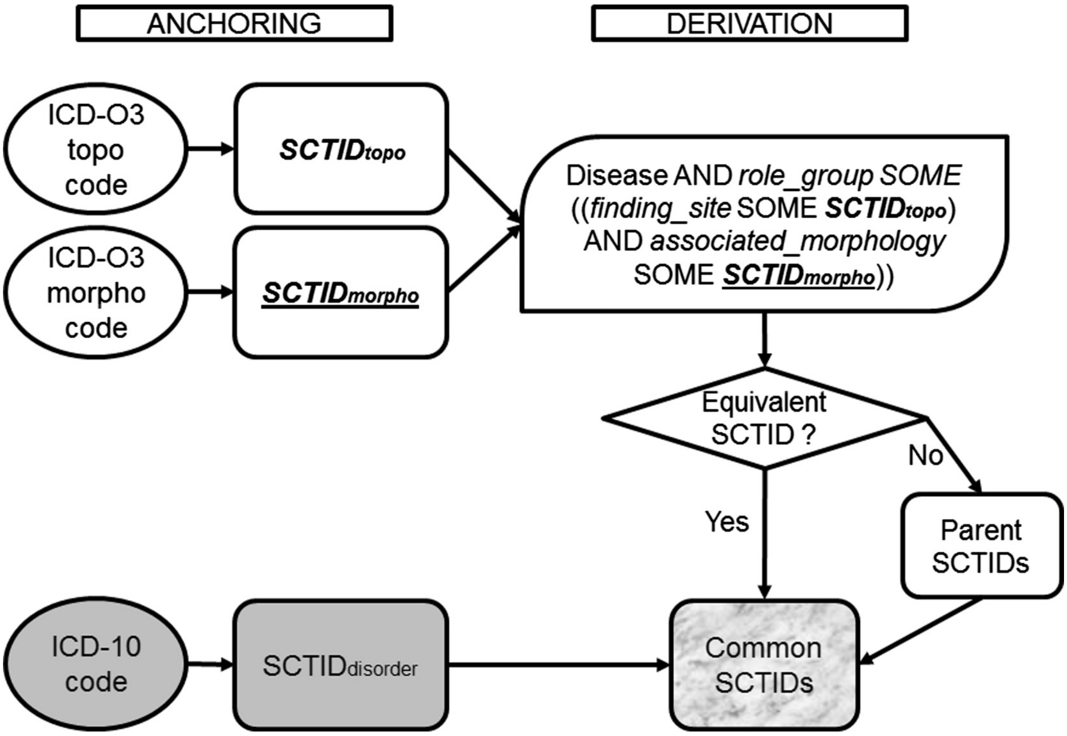


Fig. 4. The derivation stage: identifying SNOMED CT concepts of diseases that can be used as a bridge between ICD-10 and ICD-O3 concepts. For each pair of SNOMED CT concepts anchored to ICD-O3 concepts, a DL-query was performed to retrieve the expression corresponding to the disease. The equivalent concept of this DL expression was searched and if it did not exist, parent concepts were used.

executed over the inferred SNOMED CT structure obtained with the ELK reasoner. Then, either an equivalent SNOMED CT concept was found or, failing that, the parent concepts of this DL expression were recovered. We finally checked automatically if some of these SNOMED CT disease concepts were anchored to ICD-10 concepts.

4.2.2. Evaluation of the derivation stage

For the evaluation of the derivation stage, we carried out a qualitative and quantitative analyses of the integration results.

For quantitative analysis, we calculated the number of derivations found for each ICD-10 concept according to the following cardinalities:

- 1–1 derivations: an ICD-10 concept derived with a single pair of ICD-O3 morphology and topography concepts.
- 1–N derivations: an ICD-10 concept derived with more than one pair of ICD-O3 morphology and topography concepts.
- 1–0 derivations: an ICD-10 concept which could not be derived with any pair of ICD-O3 morphology and topography concepts.

We also calculated the coverage of ICD-10 and ICD-O3 concepts involved in the derivation.

For qualitative analysis, we compared our results with a gold standard, an ICD conversion file provided by the National Cancer Institute within the SEER (Surveillance, Epidemiology, and End Results) program.[2] Within this file, only the correspondences between ICD-10 and ICD-O3 concepts that participated in 1–1 anchors were used for the integration assessment. We thus calculated the overlap of our results with the 23,694 correspondences available in the SEER program conversion file.

5. Results

5.1. Anchoring stage

5.1.1. Coverage of ICD-10 and ICD-O3 concepts involved in anchors

Fig. 5 shows the distribution of ICD-10 and ICD-O3 concepts according to the resource used to establish anchors (*i.e.*, the SNCTmt or the NCI Mt). By considering the two resources (*i.e.*, anchors obtained by the SNCTmt, anchors obtained by the NCI Mt, anchors obtained by the SNCTmt and the NCI Mt), we found that more than 88.0% of ICD-10 and ICD-O3 concepts could be mapped to SNOMED CT concepts. For ICD-O3 morphology concepts, the coverage reaches 99.0% (1025/1032). It is noteworthy that for 28.0% of ICD-10 concepts, only one resource provided an anchor to a SNOMED CT concept.

5.1.2. Filtering step

Table 2 shows the impact of the filtering process steps, according to each resource used to establish the anchors. The filtering according to the hierarchy has nearly no impact on the distribution of the ICD-10 and ICD-O3 concepts in the anchors proposed by the SNCTmt. In contrast, in those recovered from the NCI Mt, the number of concepts involved in 1–N anchors decreases; a tendency which is particularly pronounced for ICD-O3 morphology concepts (from 465 to 150) and to a lesser extent for ICD-10 concepts (from 115 to 48). This diminution is accompanied by an increase in the number of ICD-O3 morphology concepts (from 539 to 847) and ICD-10 concepts (from 516 to 572) participating in 1–1 anchors. As an example, within the NCI Mt, the ICD-O3 morphology concept 9684/3-*Malignant lymphoma, immunoblastic, NOS* is anchored to the SNOMED CT concepts 109966003-*Diffuse non-Hodgkin's lymphoma, immunoblastic (disorder)* and 450909005-*Plasmablastic lymphoma (morphologic abnormality)*. The anchor between the

[2] Available at: http://seer.cancer.gov/tools/conversion/.

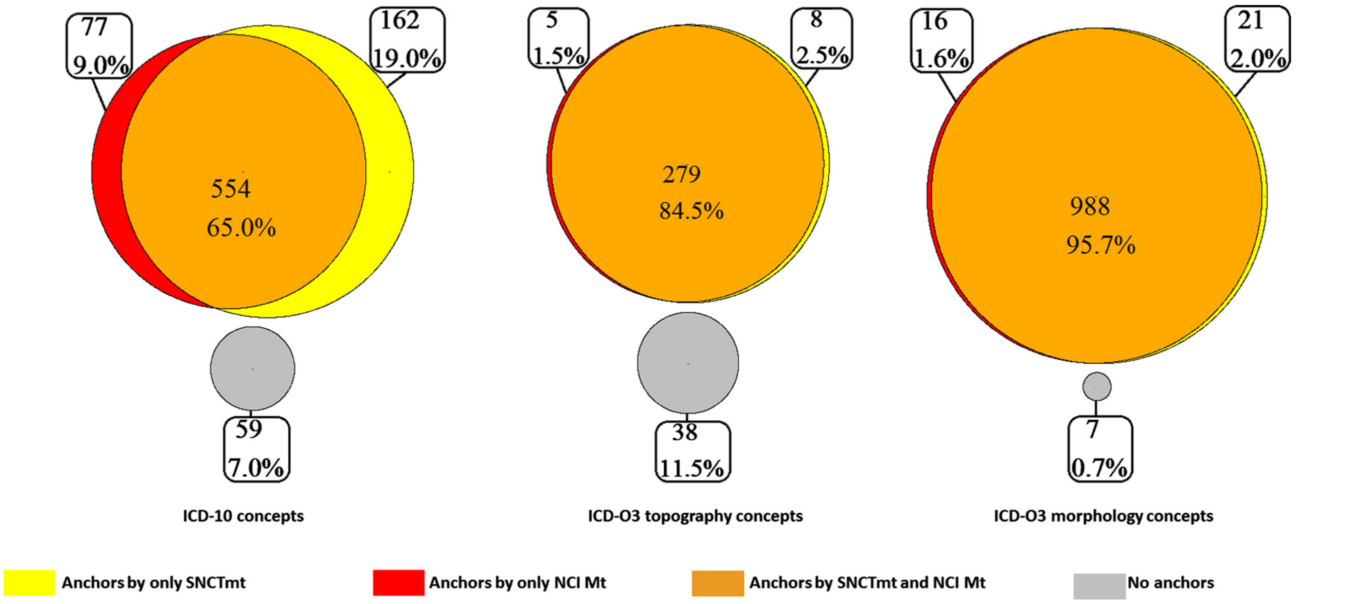


Fig. 5. Number of ICD-10 and ICD-O3 concepts involved in anchors, according to the mapping resource used to establish these anchors. The size of circles is proportional to the coverage percentage.

Table 2
Distribution of ICD-10 and ICD-O3 concepts within anchors obtained by the SNCTmt and the NCI Mt after each filtering step.

Steps	Cardinality of anchors	ICD-10		ICD-O3			
				Topography		Morphology	
		SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Initial	1–0	136	221	43	46	23	28
	1–1	79	516	4	132	960	539
	1-N	637	115	283	152	49	465
Filtering by hierarchy	1–0	136	232	44	75	24	35
	1–1	79	572	4	125	959	847
	1-N	637	48	282	130	49	150
Filtering by tumor behavior	1–0	186	537			72	91
	1–1	159	288			912	838
	1-N	507	27			48	103

ICD-O3 morphology concept (9684/3) and the concept of disease (109966003) was eliminated thanks to the filtering based on the hierarchy. The cardinality of the anchor in which this ICD-O3 concept is involved dropped from 1–N to 1–1. This step thus succeeds in reducing the number of 1–N anchors. On the other hand, some 1–1 and 1–N anchors were eliminated for 11 ICD-10 concepts, 7 ICD-O3 morphology concepts and 29 ICD-O3 topography concepts (thus resulting in additional 1–0 anchors).

The filtering according to the tumor behavior globally leads to a decrease in the number of concepts involved in 1–1 and 1–N anchors, except for ICD-10 concepts with an increasing number of 1–1 anchors coming from the SNCTmt (from 79 to 159). This step results in the elimination of many anchors, in particular for 305 ICD-10 concepts participating in anchors obtained within the NCI Mt. As an illustration, the anchor between the ICD-10 concept C47.3-*Malignant neoplasm of peripheral nerves of thorax* and the SNOMED CT concept 188325002-*Malignant neoplasm of peripheral nerve of thorax (disorder)* was deleted. According to the SNOMED CT hierarchy, this concept is described as being a tumor which can be primary or not, contrary to the ICD-10 concept which is exclusively primary. Although both concepts have the same label, they do not describe the same tumor behavior and, thus, cannot be mapped to each other.

5.1.3. Disambiguation step

The number of disambiguated concepts (*i.e.*, whose cardinality of anchors was initially 1–N and became 1–1), respectively mapped through the SNCTmt and the NCI Mt, are 289 and 14 for ICD-10 concepts, 127 and 59 for ICD-O3 topography concepts, and finally 43 and 41 for ICD-O3 morphology concepts (Table 3). An example of disambiguation is the ICD-10 concept C50.1-*Malignant neoplasm of the central portion of the breast*, which was initially mapped to the three following SNOMED CT concepts: 93745008-Primary malignant neoplasm of central portion of female breast (disorder), 708921005-*Carcinoma of central portion of breast (disorder)* and 448436006-*Sarcoma of central portion of female breast (disorder).* The disambiguation process was able to find that, among these three concepts, the concept 93745008 being the most general, it was the valid mapped concept for C50.1.

Pooled anchors of the SNCTmt and the NCI Mt resulted in the increase of ICD-10 and ICD-O3 participation in anchors. More precisely, the remaining anchors involved 706 ICD-10 concepts, 969 ICD-O3 morphology concepts and 289 ICD-O3 topography concepts. At the end of the disambiguation process, 57.2% (487/852) of ICD-10 concepts, 38.5% (127/330) of ICD-O3 topography concepts and 87.3% (901/1032) of ICD-O3 morphology concepts participated in 1–1 anchors.

5.2. Derivation stage

5.2.1. Quantitative analysis

Table 4 presents the number of ICD-10 concepts which could be derived with one or multiple pairs of ICD-O3 topography and morphology concepts and those which could not be derived at all. ICD-10 concepts were mainly derived to multiple pairs of ICD-O3 topography and morphology concepts (22.5% for 1–N derivations against 1.3% for 1–1 derivations). An example of 1–1 derivation is D13.2-*Benign neoplasm of duodenum* (ICD-10 concept) with 8850/0-*Lipoma, NOS* (ICD-O3 morphology concept) combined to C17.0-*Duodenum* (ICD-O3 topography concept). Of note, there are more 1–1 derivations between ICD-10 concepts and pairs of ICD-O3 concepts for the category "Haematological malignancy", probably because haematological tumors are very specific lesions.

Overall, by combining 1–1 and 1–N derivations, we found that 23.8% (203/852) of ICD-10 concepts could be derived with 38.5% (127/330) of ICD-O3 topography concepts and 86.0% (892/1032) of ICD-O3 morphology concepts.

5.2.2. Qualitative analysis

We found 63,142 ICD-O3 pairs which could be derived with ICD-10 concepts after the derivation process. Among them, 57,505 pairs were each derived with one ICD-10 concept and 5637 pairs were each derived with multiple ICD-10 concepts. A total of 17,474 ICD-O3 pairs were common with the 23,694 pairs described in the SEER conversion file and for 11,932 of them, our integration process found the same ICD-10 concept as the SEER conversion file. This corresponds to a recall of 0.5; a precision of 0.68 and an F-measure of 0.58. As an example, C15.9-*Esophagus, NOS* and 8504/2-*Noninfiltrating intracystic carcinoma* were derived with D00.1-*Carcinoma in situ of esophagus* both in the SEER conversion file and according to our derivation process.

For the remaining ICD-O3 pairs, our derivation process found different ICD-10 concepts than the SEER conversion file proposes. The ICD-O3 pair formed by C00.0-*External upper lip* and 8856/0-*Intramuscular lipoma* illustrates such cases. Our process resulted in derivations with D10.0-*Benign neoplasm of lip* and D17.0-*Benign lipomatous neoplasm of skin and subcutaneous tissue of head, face and neck* while the SEER conversion file describes a derivation with D17.9-*Benign lipomatous neoplasm, unspecified.*

6. Discussion

6.1. Findings

Our study consisted in integrating two biomedical terminologies that focus on diagnostic coding in the field of oncology. The difference of clinical notions represented in ICD-10 and ICD-O3 could not result in 1–1 mappings between their concepts because they are disjoint. Thus, we did not perform an alignment of these two terminologies but their integration by linking concepts through non-hierarchical relations. For this, we proposed a method for establishing appropriate semantic relations between ICD-10 pre-coordinated concepts (diseases) and ICD-O3 post-coordinated expressions (combinations of topography and morphology). Even if they describe disjoint concepts, these terminologies are organized according to a coherent structure that was used in our study for their integration. We chose SNOMED CT as a support TOR for this semantic integration not only because its domain coverage includes those of ICD-10 and ICD-O3 but also because it benefits from ontological characteristics which allowed logical inferences over its structure. Logical definitions in SNOMED CT are based on OWL-EL, explaining the choice of ELK for reasoning over its structure. OWL-EL is a "*trimmed down version of OWL that trades some expressive power for the efficiency of reasoning*". This reasoner has been shown to be sufficient to express SNOMED CT [52]. In our work, reasoning and DL-queries enabled to retrieve links that were not explicitly stated within the SNOMED-CT structure. Moreover, although built expressions based on ICD-O3 combinations could refer to anonymous classes (because not explicitly described within the SNOMED-CT), we were able to classify them and link them with an ICD-10 code.

6.1.1. Anchoring stage

By using the SNCTmt and the NCI Mt, we were able to obtain a high coverage of ICD-10 and ICD-O3 concepts within anchors. Thanks to the combined use of the two resources, we indeed found

Table 3
Disambiguation of anchors coming from the SNCTmt and the NCI Mt. 1–0 anchors do not appear because their number is not changed by the disambiguation step.

		Cardinality of anchors	ICD-10		ICD-O3			
					Topography		Morphology	
			SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Steps	Before disambiguation	1–1	159	288	4	125	912	838
		1–N	507	27	282	130	48	103
	After disambiguation	1–1	448	302	131	184	957	879
		1–N	218	13	155	71	3	62
Total			666	315	292	255	960	941

Table 4
Disambiguation of anchors coming from the SNCTmt and the NCI Mt. 1–0 anchors do not appear because their number is not changed by the disambiguation stage.

ICD-10 concepts	N*	Cardinality of ICD-10 concepts derived with pairs of ICD-O3 topography and morphology concepts						Total	
		1–0		1–1		1–N			
Benign tumor	180	57	31.6%	0	0.0%	34	18.9%	91	50.5%
Haematological malignancy	92	24	26.1%	10	10.9%	20	21.7%	54	58.7%
Unpredictable tumor	86	29	33.7%	0	0.0%	20	23.2%	49	57.0%
Tumor in situ	66	26	39.4%	1	1.5%	14	21.2%	41	62.1%
Primary malignant	388	133	34.3%	0	0.0%	99	25.5%	232	59.8%
Secondary malignant	39	14	35.9%	0	0.0%	5	12.8%	19	48.7%
Multiple tumors	1	1	100.0%	0	0.0%	0	0.0%	1	100.0%
Total	852	284	33.3%	11	1.3%	192	22.5%	487	57.1%

anchors for more than 88% of ICD-10 and ICD-O3 concepts. The highest coverage (99%) concerned ICD-O3 morphology codes, which can be explained by the fact that ICD-O3 morphology concepts were used as support for the representation of SNOMED CT histological lesions [53]. It is noteworthy that, although the overlap is important between anchors obtained by the SNCTmt and the NCI Mt, it was useful to make use of both of these resources because some anchors were found in only one of them, especially for ICD-10 concepts (19% for the SNCTmt and 9% for the NCI Mt).

The main benefit from the anchoring stage was not to create anchors but rather to improve their quality. Although ICD-10 and ICD-O3 are poorly structured, we successfully made corrections and reconciled proposed anchors, by using their structure at the filtering and disambiguation steps. These steps can thus be qualified as alignment repair processes [54]. The filtering step indeed enabled to delete anchors involving concepts that do not describe the same clinical notion. The disambiguation step managed to exclude anchors when a hierarchical relationship existed between SNOMED CT concepts involved in multiple anchors, so that only the most relevant anchor was kept. Thus, these processes highlighted and succeeded in solving the limitations of the morphosyntactic method used by the NCI Mt for establishing mappings and those of the manual method used for creating the SNCTmt. It is important to note that these two methods are the most commonly used in the literature to create mappings, like in systems described previously such as AROMA [55], ServOMap [21] and Onagui [56]. Thus, our methodology may be applied to improve the quality of mappings created by any such application. Indeed, our method is independent of strategies used for creating mappings, because it is only based on the structure of SNOMED CT, ICD-10 and ICD-O3.

6.1.2. Derivation stage

6.1.2.1. Derivation strategy. In the derivation process, we looked for equivalent, and parent if necessary, concepts of the DL expression corresponding to a pair of ICD-O3 concepts. ICD-10 represents nosologic entities and an ICD-10 concept can represent one or more entities. The notions represented by a combination of ICD-O3 concepts may correspond exactly to the nosologic entity represented by the ICD-10 concept, in which case an equivalence can be found. In contrast, the ICD-O3 combination may represent a nosologic entity which is part of a group of entities represented by an ICD-10 concept. In this situation, subsumption relations are thus of interest.

6.1.2.2. Derivation coverage. We were able to derive 86% of the ICD-O3 morphology concepts, 36% of the ICD-O3 topography concepts and 24% of the ICD-10 concepts. The coverage of ICD-10 concepts is correlated with the coverage of ICD-03 topography concepts because ICD-10 concepts related to cancer diagnoses are grouped according to the anatomical localization of the tumor. Thus, the absence of anchors for a given ICD-O3 topography automatically implies the absence of anchors for the ICD-10 concepts involving this anatomical localization. Conversely, the coverage of ICD-O3 morphology concepts is high. This can be explained by the facts that: (i) the same histological lesion may exist for different anatomical localizations, and (ii) the description of histological lesions in ICD-O3 is more precise than in ICD-10. This difference in the level of precision also explains the numerous 1–N derivations.

6.1.2.3. Derivation quality. The derivation stage enabled to find an ICD-10 concept for 74% (17,474/23,694) of ICD-O3 pairs of the SEER conversion file. Moreover, our integration process correctly and automatically generated 50% of the correspondences between an ICD-10 concept and a pair of ICD-O3 concepts described in the SEER conversion file.

A potential explanation of the divergences observed between our derivation process and correspondences proposed by the SEER is that its conversion file is based on rules of cancer registries. Conversely, our derivation process intends to relate ICD-O3 combinations to ICD-10 concepts based on their semantics. As a result, our process can find multiple derivations for a single combination whereas the SEER proposes only one of them. For instance, in the SEER conversion file and according to our derivation process, the ICD-O3 pair C75.3-*Pineal gland* and 9769/1-*Immunoglobulin deposition disease* was integrated with D47.9-*Neoplasm of uncertain or unknown behavior of lymphoid, haematopoietic and related tissue, unspecified* (according to the rule 4.1 of cancer registries for recording an haematopoietic disease [57]). However, our derivation process also proposed D44.5-*Neoplasm of uncertain behavior of pineal gland* for this pair. Although the later derivation is significant, it has not been retained by the SEER. This finding highlights that our process does not depend on specific conversion rules, but only on the semantics provided by SNOMED-CT.

Another consequence of our process was the derivation of pairs that are not medically relevant. An example of such irrelevant combinations is the ICD-O3 pair C50.2-*Upper-inner quadrant of breast* and 8153/1-*Gastrinoma, NOS* which was integrated with the ICD-10 concept D48.6-*Neoplasm of uncertain or unknown behavior of Breast*. Indeed, "gastrinoma" is a specific morphologic abnormality of the digestive tract so this tumor cannot appear with breast as a primary site. Confronting derivation results with cancer registries' data is a perspective that would make it possible to keep only the ICD-O3 pairs which are effectively used in practice to record health data. However, it is necessary to underline that our derivation process enable to take into account the imperfect but informative coding that may exist in real data (*i.e.*, coding error).

6.2. Integration process and evaluation limitations

Our integration of ICD-10 and ICD-O3 concepts remains incomplete. The main limitation of our methods concerns 1–N and 1–0 anchors, which were not derived. For 1–N anchors, the disambiguation process needs to be improved. Some 1–N concepts were still present after the disambiguation of results obtained by the SNCTmt and the NCI Mt but others were also created when pooling anchors coming from these two resources. In many cases, this is a consequence of concepts that are incorrectly represented as siblings within the SNOMED CT structure. As an example, the ICD-O3 morphology concept 8831/0-*Histiocytoma, NOS* was involved in a 1–1 anchor with 128741006-*Deep histiocytoma (morphologic abnormality)* according to the SNCTmt and in a 1–1 anchor with 302843004-*Histiocytoma (morphologic abnormality)* according to the NCI Mt. By pooling the anchors of the two resources, 8831/0-*Histiocytoma, NOS* finally participated in a 1–N anchor because the two concepts 128741006-*Deep histiocytoma (morphologic abnormality)* and 302843004-*Histiocytoma (morphologic abnormality)* are erroneously described as siblings in SNOMED CT. Indeed, they must clearly be related through a subsumption relationship. Other hierarchical and non-hierarchical semantic links must be sought by the disambiguation process because SNOMED CT apparently does not contain appropriate links between some of its concepts. Therefore, a potential strategy for improving the disambiguation process would be to search for such semantic links in other knowledge resources. As an example, the Foundational Model of Anatomy (FMA) [58] may be a good candidate to identify relations between SNOMED CT concepts which are anchored to a given ICD-O3 topography concept.

6.3. Comparison with previous works

The most similar previous work compared to our study is the one realized by Jouhet et al. [38], who also tried to integrate ICD-10 and ICD-O3 thanks to a support TOR, namely the NCI thesaurus.

If we compare their results with ours, we derived 888 ICD-O3 morphology concepts against 860 for them. By contrast, as our model only considers 1–1 anchors for the derivation stage, the high proportion of 1–N anchors for ICD-O3 topography concepts leads to a lower coverage of ICD-O3 topography concepts (127) and ICD-10 concepts (203) involved in derivations compared to the coverage obtained by Jouhet et al., being respectively 278 and 302. Thus, one of our prospects is the fusion of our results with those obtained by Jouhet et al. We would like to check if the use of the NCI thesaurus could improve our semantic integration. In particular, for concepts having no anchors with any SNOMED CT concept, such concepts could be mapped to NCI thesaurus concepts. Finally, we believe that merging the results of both studies will highlight complementarities of the NCI thesaurus and the SNOMED CT. We chose SNOMED CT because it has been used to support the semantic integration of various biomedical TORs in previous works. For instance, Brown et al. [59] used it to align the Veterans Benefits Administration (VBA) disability code set and the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9CM). In this work, authors used a morphosyntactic method and compared two approaches. The first one was a direct alignment of VBA and ICD-9CM without using any other resource, while the second approach used the SNOMED CT as a support TOR. The use of SNOMED CT increased the alignment coverage, which illustrates that SNOMED CT is able to cover various clinical domains of medicine. This work differs from ours in that authors did not use the structure of the SNOMED CT to find semantic links between disjoint concepts or to ensure the quality of the mapping process. Finally, the repair process performed in this study was made manually by domain experts in a consensual way. Another example is the work of Bakhshi-Raiez et al. [60], who used SNOMED CT to align APACHE II and APACHE IV, which are two versions of a classification system used to encode the reasons for intensive care admission. Firstly, authors manually created mappings between SNOMED CT concepts and those of APACHE II and IV. Then, authors used the SNOMED CT structure to retrieve SNOMED CT concepts which had hierarchical links (especially the *part_of* relationship) with already mapped concepts. Thus, the common SNOMED CT concepts mapped to APACHE II and IV concepts constituted the bridges between the two classifications. As in our study, authors used the structure of SNOMED CT to establish anchors between two TORs but they did not have to realize a semantic integration of disjoint concepts. A challenge raised by the integration of ICD-10 and ICD-O3 was to align pre-coordinated concepts with post-coordinated expressions. To address this issue, Dhombres et al. [26] have implemented a strategy requiring that one of the two TORs to be aligned must have sophisticated labels and the other one must be able to carry out post-coordination. Because ICD-10 and ICD-O3 do not have these two characteristics, we could not follow such a strategy and we had to propose an alternative one.

6.4. Integration process advantages

An analytical reading of our semantic integration process gives the possibility to understand and correct our methodological choices but also to indirectly observe limitations in the structure of SNOMED CT, ICD-10 and ICD-O3. Indeed, the anchoring and derivation stages are based on the SNOMED CT structure, and in particular, on the subsumption relations existing within SNOMED CT. Our method enabled to identify some limitations and specificities of the SNOMED CT structure that were already described in the literature, such as the "*absence of difference in the description between children and parents*" [61,41]. For example, SNOMED CT does not consider 109271004-*Melanocytic nevus of lip (disorder)* as being a benign tumor, erroneously. Although being correct, its anchor with the ICD-10 concept D22.0-*Melanocytic naevi of lip* was deleted during the filtering step (by tumor behavior). Another example is the anchor of the ICD-O3 morphology concept 8151/3-*Insulinoma, malignant* with the SNOMED CT concept 20955008-*Insulinoma, malignant (morphologic abnormality)*. Although this anchor is correct, it was erroneously derived with some ICD-10 concepts of benign tumors because the SNOMED CT concept 20955008 is a descendant of 3898006-*Neoplasm, benign (morphologic abnormality)* and 86049000-*Malignant neoplasm, primary (morphologic abnormality)*. This inconsistency illustrates an uncontrolled use of the subsumption relationship in SNOMED CT, which is called *is_a* overloading [41].

Other TOR-related problems were encountered during the semantic integration process. Indeed, from the beginning of the process, we identified concepts that did not participate in anchors. The concepts which could not be mapped are mainly ICD-O3 topography concepts with codes (.8) describing an overlapping anatomical site (*e.g.*, C63.8-*Overlapping lesion of male genital organs*), as well as ICD-10 and ICD-O3 concepts which use the category "other" for unlisted diagnoses or histological lesions. The ICD-10 concept C45.7-*Mesothelioma of other sites* is such an example. ICD-10 enumerates three anatomical sites for mesothelioma (C45.0-Mesothelioma of pleura, C45.1-*Mesothelioma of peritoneum* and C45.2-*Mesothelioma of pericardium*), and C45.7 encodes for all mesothelioma that are not pleura, peritoneum and pericardium mesothelioma [7,9]. This representation is made because of the epidemiologic objectives of ICD-10 and ICD-O3. The objectives of SNOMED CT being different, it does not include such concepts. To address this issue, we could look for structural proximities between the concepts belonging to the "other" category and concepts already anchored, like ServOMap [21] and SAMBO [18] do. For these particular concepts, we could indeed search for their parent concepts having anchors with a SNOMED CT concept and some of the direct descendants of this SNOMED CT concept, which are not already anchored, could be mapped to the ICD-10/ ICD-O3 concept belonging to the "other" category.

7. Conclusion

Our study set up a model for integrating ICD-10 pre-coordinated concepts and ICD-O3 post-coordinated expressions. This integration was based on a support TOR, SNOMED CT, which describes the semantic relations existing between clinical notions represented in ICD-10 and ICD-O3. Our methods constitute a repair process, which can be used by systems creating mappings based on manual and morphosyntactic approaches. We also indirectly conducted an audit of SNOMED CT, ICD-10 and ICD-O3. The semantic integration process may be improved, especially by taking into account the specificities of used TORs, by using other support TORs and by combining our results with those obtained in previous works.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.08.013.

References

[1] C. Safran, M. Bloomrosen, W.E. Hammond, S. Labkoff, S. Markel-Fox, P.C. Tang, D.E. Detmer, Toward a national framework for the secondary use of health data: an american medical informatics association white paper, J. Am. Med. Inform. Assoc. 14 (1) (2007) 1–9, http://dx.doi.org/10.1197/jamia.M2273. <http://jamia.oxfordjournals.org/cgi/doi/10.1197/jamia.M2273>.

[2] M. TB, D. AS, The inevitable application of big data to health care, JAMA 309 (13) (2013) 1351–1352, http://dx.doi.org/10.1001/jama.2013.393. Available from: arXiv:/data/journals/jama/926712/jvp130007_1351_1352.pd.

[3] E.N. MacKay, A.H. Sellers, The Ontario cancer incidence survey, 1964–1966: a new approach to cancer data acquisition, Can. Med. Assoc. J. 109(6) (1973) 489 passim.

[4] O.M. Jensen, D.M. Parkin, International Association of Cancer Registries, International Agency for Research on Cancer (Eds.), Cancer Registration: Principles and Methods, No. 95 in IARC Scientific Publications, Internat. Agency for Research on Cancer [u.a.], Lyon, 1991, oCLC: 24293858.

[5] D.M. Parkin, The role of cancer registries in cancer control, Int. J. Clin. Oncol. 13 (2) (2008) 102–111, http://dx.doi.org/10.1007/s10147-008-0762-6. <http://link.springer.com/10.1007/s10147-008-0762-6>.

[6] V. Jouhet, G. Defossez, Automated selection of relevant information for notification of incident cancer cases within a multisource cancer registry:, Meth. Inform. Med. 52 (5) (2013) 411–421, http://dx.doi.org/10.3414/ME12-01-0101. CRISAP, CoRIM, P. Ingrand <http://www.schattauer.de/index.php?id=1214&doi=10.3414/ME12-01-0101>.

[7] A. Fritz, C. Percy, K. Shanmugaratnam, L. Sobin, D.M. Parkin, S. Whelan, International classification of diseases for oncology: ICD-O, third ed., World Health Organization, Geneva, 2000, oCLC: 248314653.

[8] C. Colin, R. Ecochard, F. Delahaye, G. Landrivon, P. Messy, E. Morgon, Y. Matillon, Data quality in a DRG-based information system, Int. J. Qual. Health Care 6 (3) (1994) 275–280, http://dx.doi.org/10.1093/intqhc/6.3.275. <http://intqhc.oxfordjournals.org/cgi/doi/10.1093/intqhc/6.3.275>.

[9] W.H. Organization, International Statistical Classification of Diseases and Related Health Problems 10th revision, 2010th Edition, vol. 2, 2011. <http://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf>.

[10] P. Zweigenbaum, Encoder linformation mdicale: des terminologies aux systmes de reprsentation des connaissances, Innov. Stratgique Inform. Sant 2 (1999) 5. <http://perso.limsi.fr/pz/FTPapiers/ZweigenbaumISIS99.pdf.g>.

[11] R. Studer, V. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data Knowl. Eng. 25 (1-2) (1998) 161–197, http://dx.doi.org/10.1016/S0169-023X(97)00056-6. <http://linkinghub.elsevier.com/retrieve/pii/S0169023X97000566>.

[12] J. Euzenat, P. Shvaiko, Ontology Matching: With 67 Figures and 18 Tables, Springer, Berlin, 2007.

[13] H. Saitwal, D. Qing, S. Jones, E.V. Bernstam, C.G. Chute, T.R. Johnson, Cross-terminology mapping challenges: a demonstration using medication terminological systems, J. Biomed. Inform. 45 (4) (2012) 613–625, http://dx.doi.org/10.1016/j.jbi.2012.06.005.

[14] K. Giannangelo, J. Millar, Mapping SNOMED CT to ICD-10, Stud. Health Technol. Inform. 180 (2012) 83–87.

[15] J. Souvignet, J.-M. Rodrigues, Toward a Patient Safety Upper Level Ontology, Stud. Health Technol. Inform. (2015) 160–164, http://dx.doi.org/10.3233/978-1-61499-512-8-160. <http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-511-1&spage=160&doi=10.3233/978-1-61499-512-8-160>.

[16] M. Cheatham, Z. Dragisic, J. Euzenat, D. Faria, A. Ferrara, G. Flouris, I. Fundulaki, R. Granada, V. Ivanova, E. Jimnez-Ruiz, others, Results of the ontology alignment evaluation initiative 2015, in: 10th ISWC Workshop on Ontology Matching (OM), No Commercial Editor, 2015, pp. 60–115. <https://hal.archives-ouvertes.fr/hal-01254907/>.

[17] P. Shvaiko, J. Euzenat, Ontology matching: state of the art and future challenges, IEEE Trans. Knowl. Data Eng. 25 (1) (2013) 158–176, http://dx.doi.org/10.1109/TKDE.2011.253. <http://ieeexplore.ieee.org/document/6104044/>.

[18] H. Tan, P. Lambrix, SAMBO results for the ontology alignment evaluation initiative 2007, in: Proceedings of the 2nd International Conference on Ontology Matching-Volume 304, CEUR-WS. org, Springer, 2007, pp. 236–243. <http://disi.unitn.it/p2p/OM-2007/12-o-SAMBO.pdf>.

[19] A. Tian, J.F. Sequeda, D.P. Miranker, Qodi: Query as context in automatic data integration, in: International Semantic Web Conference, Springer, 2013, pp. 624–639. <http://link.springer.com/chapter/10.1007/978-3-642-41335-3_39>.

[20] I.F. Cruz, C. Stroe, F. Caimi, A. Fabiani, C. Pesquita, F.M. Couto, M. Palmonari, Using AgreementMaker to align ontologies for OAEI 2011, in: Proceedings of the 6th International Conference on Ontology Matching-Volume 814, CEUR-WS. org, 2011, pp. 114–121. <http://dl.acm.org/citation.cfm?id=2887550>.

[21] G. Diallo, An effective method of large scale ontology matching, J. Biomed. Seman. 5 (1) (2014) 44, http://dx.doi.org/10.1186/2041-1480-5-44.

[22] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (11) (1995) 39–41. <http://dl.acm.org/citation.cfm?id=21974>.

[23] C.J. Mungall, C. Torniai, G.V. Gkoutos, S.E. Lewis, M.A. Haendel, Uberon, an integrative multi-species anatomy ontology, Genome Biol. 13 (1) (2012) 1. <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2012-13-1-r>.

[24] A. Rector, L. Iannone, Lexically suggest, logically define: Quality assurance of the use of qualifiers and expected results of post-coordination in SNOMED CT, J. Biomed. Inform. 45 (2) (2012) 199–209, http://dx.doi.org/10.1016/j.jbi.2011.10.002. <http://linkinghub.elsevier.com/retrieve/pii/S1532046411001687>.

[25] R.H. Dolin, K.A. Spackman, D. Markwell, Selective retrieval of pre- and post-coordinated SNOMED concepts, in: Proceedings AMIA Symposium, 2002, pp. 210–214.

[26] F. Dhombres, R. Winnenburg, T. Case James, O. Bodenreider, Extending the coverage of phenotypes in SNOMED CT through post-coordination, Stud. Health Technol. Inform. (2015) 795–799, http://dx.doi.org/10.3233/978-1-61499-564-7-795. <http://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-563-0&spage=795&doi=10.3233/978-1-61499-564-7-795>.

[27] Y. Kalfoglou, M. Schorlemmer, IF-Map: an ontology-mapping method based on information-flow theory, in: Journal on Data Semantics I, Springer, 2003, pp. 98–127. <http://link.springer.com/chapter/10.1007/978-3-540-39733-5_5>.

[28] K. Tatane, B. Er-Raha, C.-e. Cherkaoui, S. Mouhim, Alignment methodological approach of evolving domain sub-ontologies using terminological and structural matchers applied to tourism domain, Int. J. Comput. Appl. 123 (15) (2015) 6–13, http://dx.doi.org/10.5120/ijca2015905710. <http://www.ijcaonline.org/research/volume123/number15/tatane-2015-ijca-905710.pdf>.

[29] C. Pinkel, C. Binnig, P. Haase, C. Martin, K. Sengupta, J. Trame, How to best find a partner? An evaluation of editing approaches to construct R2rml mappings, in: European Semantic Web Conference, Springer, 2014, pp. 675–690. <http://link.springer.com/chapter/10.1007/978-3-319-07443-6_45>.

[30] B. Safar, C. Reynaud, F. Calvier, Techniques dalignement dontologies bases sur la structure dune ressource complmentaire, 1res J. Francophones Ontol. (JFO 2007) (2007) 21–35. <https://www.lri.fr/perso/cr/papiers/2007/JFO07.pdf>.

[31] S. Abburu, A survey on ontology reasoners and comparison, Int. J. Comput. Appl. 57(17). <http://search.proquest.com/openview/cfaf4b8c27eb7c9d8210 6b2f5aed968c/1?pq-origsite=gscholar>.

[32] H. Stuckenschmidt, F. Van Harmelen, L. Serafini, P. Bouquet, F. Giunchiglia, Using C-OWL for the Alignment and Merging of Medical Ontologies. <http://eprints.biblio.unitn.it/523/>.

[33] F. Mougin, M. Dupuch, N. Grabar, Improving the mapping between MedDRA and SNOMED CT, in: Artificial Intelligence in Medicine, Springer, 2011, pp. 220–224. <http://link.springer.com/10.1007%2F978-3-642-22218-4_27>.

[34] T.Y. Kim, A. Coenen, N. Hardiker, Semantic mappings and locality of nursing diagnostic concepts in UMLS, J. Biomed. Inform. 45 (1) (2012) 93–100, http://dx.doi.org/10.1016/j.jbi.2011.09.002. <http://linkinghub.elsevier.com/retrieve/pii/S1532046411001626>.

[35] R. Winnenburg, L. Rodriguez, F.M. Callaghan, A. Sorbello, A. Szarfman, O. Bodenreider, aligning pharmacologic classes between MeSH and ATC, in: VDOS + DO@ ICBO, 2013. <http://ceur-ws.org/Vol-1061/Paper5_vdos2013.pdf>.

[36] A. Burgun, O. Bodenreider, Issues in integrating epidemiology and research information in oncology: experience with ICD-O3 and the NCI Thesaurus, in: AMIA Annu. Symp. Proc., 2007, pp. 85–89. <http://morc2.nlm.nih.gov/pubs/pdf/2007-amia-ab.pdf>.

[37] G. Jiang, H.R. Solbrig, C.G. Chute, Quality evaluation of cancer study common data elements using the UMLS semantic network, J. Biomed. Inform. 44 (2011) S78–S85, http://dx.doi.org/10.1016/j.jbi.2011.08.001. <http://linkinghub.elsevier.com/retrieve/pii/S1532046411001286>.

[38] V. Jouhet, F. Mougin, B. Brchat, F. Thiessard, Building a model for disease classification integration in oncology, an approach based on the national cancer institute thesaurus, J. Biomed. Semant. 8(1). doi:http://dx.doi.org/10.1186/s13326-017-0114-4. <http://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-017-0114-4>.

[39] G. Hja, G. Surjn, P. Varga, Ontological analysis of SNOMED CT, BMC Med. Inform. Decis. Mak. 8 (Suppl 1) (2008) S8, http://dx.doi.org/10.1186/1472-6947-8-S1-S8. <http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-8-S1-S8>.

[40] IHTSDO, SNOMED CT Starter Guide, February 2014. <http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_StarterGuide_Current-en-US_INT_20140222.pdf>.

[41] O. Bodenreider, B. Smith, A. Kumar, A. Burgun, Investigating subsumption in DL-based terminologies: a Case Study in SNOMED CT, in: KR-MED, vol. 2004, 2004, pp. 12–20. <http://mor.nlm.nih.gov/pubs/pdf/2004-krmed-ob.pdf>.

[42] IHTSDO, SNOMED CT Technical implementation Guide January 2015 International Release (GB English), 2015. <http://ihtsdo.org/fileadmin/user_upload/doc/download/doc_TechnicalImplementationGuide_Current-en-GB_INT_20150131.pdf?ok>.

[43] O. Bodenreider, Oncology in SNOMED CT, May 2015. <https://mor.nlm.nih.gov/pubs/pres/20150513-CancerBigData.pdf>.

[44] K.A. Spackman, R. Dionne, E. Mays, J. Weis, Role grouping as an extension to the description logic of Ontylog, motivated by concept modeling in SNOMED, in: Proceedings of the AMIA Symposium, American Medical Informatics Association, Springer, 2002, p. 712. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2244464/>.

[45] R. Cornet, S. Schulz, Relationship groups in SNOMED CT, Stud. Health Technol. Inform. 150 (2009) 223–227.

[46] M. Horridge, N. Drummond, J. Goodwin, A.L. Rector, R. Stevens, H. Wang, The manchester OWL syntax, OWLED, vol. 216, 2006.

[47] IHTSDO, Mapping SNOMED CT to ICD-10 Technical Specifications, January 2015.

[48] NCI Metathesaurus, 2016. <https://ncimeta.nci.nih.gov/ncimbrowser/>.

[49] P.L. Schuyler, W.T. Hole, M.S. Tuttle, D.D. Sherertz, The UMLS Metathesaurus: representing different views of biomedical concepts, Bull. Med. Library Assoc. 81 (2) (1993) 217. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC225764>.

[50] Y. Kazakov, M. Krtzsch, F. Simancik, ELK Reasoner: Architecture and Evaluation, in: ORE, 2012. <https://www.uni-ulm.de/fileadmin/website_uni_ulm/iui.inst.090/Publikationen/2012/KazKroSim12ELK_ORE.pdf>.

[51] K. Dentler, R. Cornet, Intra-axiom redundancies in SNOMED CT, Artif. Intell. Med. 65 (1) (2015) 29–34, http://dx.doi.org/10.1016/j.artmed.2014.10.003.

[52] OWL 2 Web Ontology Language Profiles, second edition. <https://www.w3.org/TR/owl2-profiles/#OWL_2_EL>.

[53] IHTSDO, SNOMED CT Editorial Guide, 2016. <https://confluence.ihtsdotools.org/display/DOCEG/SNOMED+CT+Editorial+Guide>.

[54] C. Pesquita, D. Faria, E. Santos, F.M. Couto, To repair or not to repair: reconciling correctness and coherence in ontology reference alignments, in: Proceedings of the 8th International Conference on Ontology Matching, vol. 1111, CEUR-WS. org, 2013, pp. 13–24. <http://dl.acm.org/citation.cfm?id=2874495>.

[55] J. David, F. Guillet, H. Briand, Matching directories and OWL ontologies with AROMA, in: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, ACM, 2006, pp. 830–831. <http://exmo.inrialpes.fr/jdavid/publies/JDavid_CIKM_2006.pdf>.

[56] OnAGUI - Ontology Alignment GUI download | SourceForge.net. <https://sourceforge.net/projects/onagui/>.

[57] M. Curado, N. Okamoto, L. Ries, H. Sriplung, J. Young, M. Carli, I. Izarzugaza, B. Koscianska, E. Demaret, J. Ferlay, M. Parkin, J. Tyczynski, S. Whelan, International Rules for Multiple Primary Cancers (ICD-O, third ed.), 2004.

[58] L.T. Detwiler, J.L. Mejino, J.F. Brinkley, From frames to OWL2: converting the foundational model of anatomy, Artif. Intell. Med. 69 (2016) 12–21, http://dx.doi.org/10.1016/j.artmed.2016.04.003. <http://linkinghub.elsevier.com/retrieve/pii/S093336571630152X>.

[59] S.H. Brown, C.S. Husser, D. Wahner-Roedler, S. Bailey, L. Nugent, K. Porter, B.A. Bauer, P.L. Elkin, Using SNOMED CT as a reference terminology to cross map two highly pre-coordinated classification systems, Stud. Health Technol. Inform. 129 (Pt 1) (2007) 636–639.

[60] F. Bakhshi-Raiez, R. Cornet, R.J. Bosman, H. Joore, N.F. de Keizer, Using SNOMED CT to identify a crossmap between two classification systems: a comparison with an expert-based and a data-driven strategy, Stud. Health Technol. Inform. 160 (Pt 2) (2010) 1035–1039.

[61] S. Schulz, B. Suntisrivaraporn, F. Baader, M. Boeker, SNOMED reaching its adolescence: ontologists and logicians health check, Int. J. Med. Inform. 78 (2009) S86–S94, http://dx.doi.org/10.1016/j.ijmedinf.2008.06.004. <http://linkinghub.elsevier.com/retrieve/pii/S1386505608000919>.

Appendix C

Original ICBO article: Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability

Comparing the representation of medicinal products in RxNorm and SNOMED CT – Consequences on interoperability

Jean Noel Nikiema[a], Olivier Bodenreider[b]

[a] Bordeaux Population Health Research Center, ERIAS, Univ. Bordeaux, Inserm UMR 1219, F-33000, Bordeaux, France
[b] U.S. National Library of Medicine National Institutes of Health Bethesda, Maryland, USA

Abstract

Objectives: To compare the representation of medicinal products in RxNorm and SNOMED CT and assess the consequences on interoperability. Methods: To compare the two models, we manually establish equivalences between the types and definitional features of medicinal products entities in RxNorm and SNOMED CT. We highlight their similarities and differences. Results: Both models share major definitional features including ingredient (or substance), strength and dose form. SNOMED CT is more rigorous and better aligned with international standards. In contrast, RxNorm contains implicit knowledge, simplifications and ambiguities, but its model is simpler. Conclusions: Since their models are largely compatible, medicinal products from RxNorm and SNOMED CT are expected to be interoperable. However, specific aspects of the alignment between the two models require particular attention.

Keywords:

RxNorm; SNOMED CT; medicinal products.

Background

Drug terminologies, such as RxNorm and the medicinal product hierarchy of SNOMED CT (Systematized Nomenclature of Medicine-Clinical Terms), support multiple use cases, including electronic prescription, drug information exchange, medication reconciliation, and analytics (including pharmacovigilance) (1,2). A formal representation of medicinal products is needed for the principled development and maintenance of such drug terminologies, as well as for precisely aligning existing drug terminologies (3).

Many definitional characteristics of medicinal products are similar among drug terminologies. For example, clinical drugs are generally defined in terms of ingredient, strength and dose form. However, the level of formality and the formalism used for representing medicinal products may differ among terminologies. Some attributes may also be specific to some terminologies (especially for country-dependent attributes, such as packaging information).

In addition to existing drug terminologies, international standards have been developed for the representation of medicinal products, such as IDMP (Identification of Medicinal Products). IDMP (4), a collection of recommendations from the International Standards Organization (ISO).

Interoperability among drug terminologies is especially important for exchanging drug information internationally. For example, a medication list established with RxNorm in the U.S. could be made available to any electronic health record (EHR) system in the world, in which drugs are represented using SNOMED CT. To fully support this use case, however, the models of medicinal products in RxNorm and SNOMED CT must be compatible, such that one can be accurately translated into the other.

We focus on RxNorm and SNOMED CT, because RxNorm is the standard drug terminology in the U.S. and SNOMED CT is the largest clinical terminology in the world, supported by a consortium of over 40 countries. While the RxNorm model has been analyzed (5,6), and reused to create others standards (7,8) and to integrate drug terminologies worldwide (8), there has not been a detailed comparison between RxNorm and SNOMED CT. Moreover, the SNOMED CT model for medicinal products is particularly interesting, because it was recently updated, in part to comply with IDMP requirements (9).

In this investigation, we compare the representation of medicinal products in RxNorm and SNOMED CT. The objective of our work is to analyze their similarities and differences and the consequences of these differences on interoperability between the two terminologies.

Methods and results

In this section, we describe the models of RxNorm and SNOMED CT with focus on their definitional characteristics. Then we identify similarities and differences between the two models.

The SNOMED CT model for medicinal products

The SNOMED CT, the largest clinical terminology in the world, is an international clinical terminology based on a formal concept model (10). SNOMED CT recently published a new model for the representation of medicinal products integrating requirements from IDMP (9). The model was developed to support international usage. Therefore, it is restricted to generic drugs and does not represent packaging information or branded drugs, which tend to be country-specific.

In accordance with requirements from IDMP, clinical drugs are represented in a closed worldview. This means that characteristics used to define clinical drugs must be sufficient and what is not stated is false. In contrast, in the open worldview, what is not stated is potentially true. For example, the representation of a clinical drug containing Atorvastatin must clearly state that this product only contains the substance Atorvastatin as its active ingredient (i.e., without any other active ingredient). In the open worldview, products containing Atorvastatin could also contain other active ingredients, e.g., Amlodipine.

As shown in Figure 1, the representation of medicinal products in SNOMED CT is based on a model with six (6) entities, arranged in a subclass hierarchy:

- Two ***medicinal product*** entities, in open and closed worldview (e.g., open worldview: *108655000 | Product containing cetirizine (medicinal product)* and closed worldview: *775140005 | Product containing only cetirizine (medicinal product)*).
- Two ***medicinal product form*** entities, in open and closed worldview, (e.g., open worldview: *768065006 | Product containing cetirizine in oral dose form (medicinal product form)* and closed worldview: *778701007 | Product containing only cetirizine in oral dose form (medicinal product form)*).
- One ***medicinal product precisely*** entity in closed worldview only (optional entity, currently not represented in SNOMED CT – hypothetical example: *Product containing only cetirizine hydrochloride (medicinal product)*).
- One ***clinical drug*** entity, in closed worldview only (e.g., *320818006 | Product containing precisely cetirizine hydrochloride 10 milligram/1 each conventional release oral tablet (clinical drug)*).

The representation of SNOMED CT entities is based on "definitional roles" and related "types of values" in SNOMED CT (Figure 1):

- **Substance** is the type of values for the *active ingredient*, *precise active ingredient* and *basis of strength* roles, for example *372523007 | Cetirizine (substance)* and *108656004 | Cetirizine hydrochloride (substance)*. (The basis of strength is the substance in reference to which strength is defined.)
- **Unit of measure** is the type of values for the *strength unit* roles, for example, *258684004 | milligram (qualifier value)*.
- **Number** is the type of values for the *strength value* roles, for example, *3445001 | 10 (qualifier value)*.
- **Pharmaceutical dose form** is the type of values for the *manufactured dose form* role, for example, *421026006 | Conventional release oral tablet (dose form)*.
- **Unit of presentation** is the type of values for the *unit of presentation* role, for example, *732936001 | Tablet (unit of presentation)*.

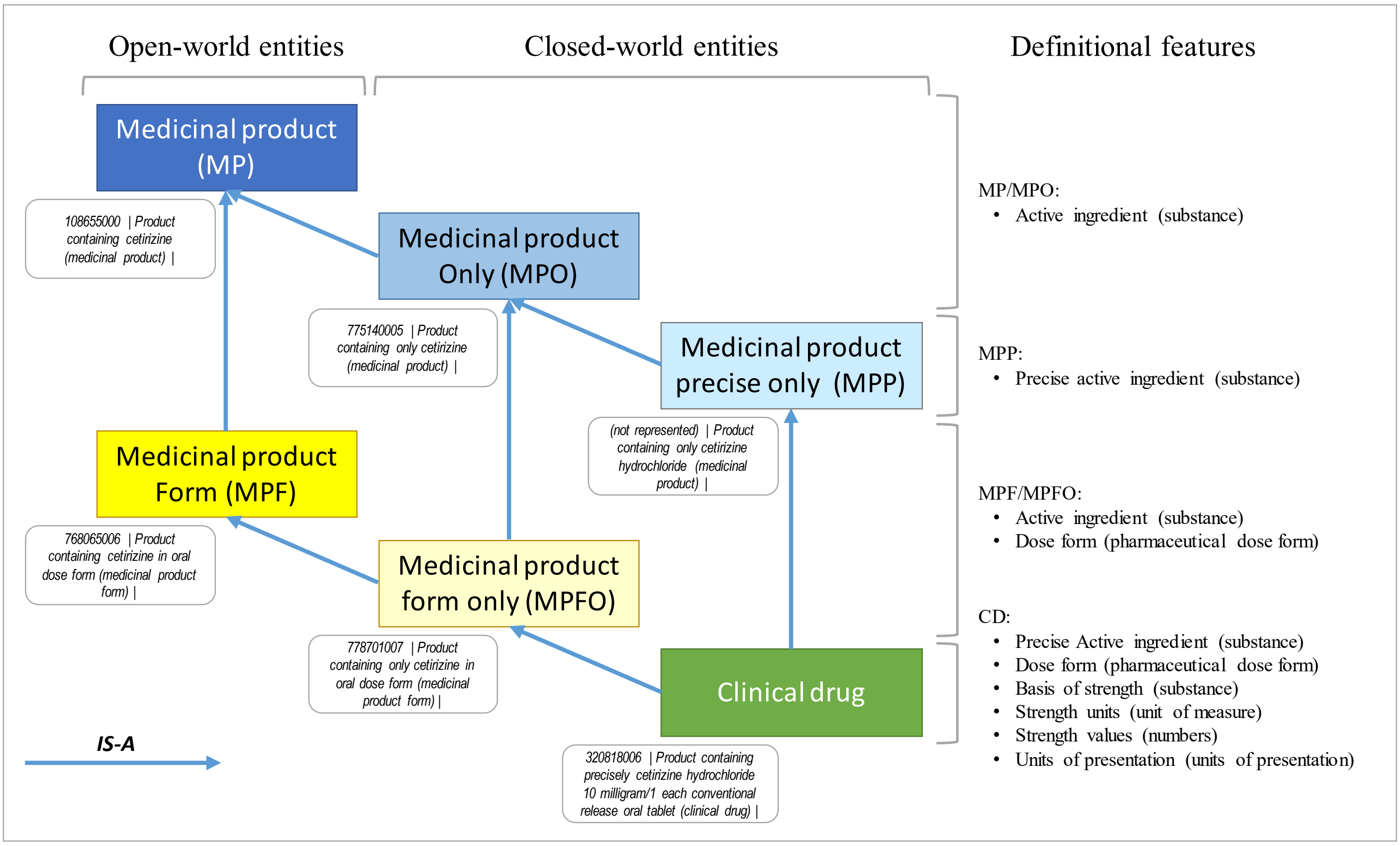

Open-world entities
Closed-world entities
Definitional features
Medicinal product (MP)
108655000 | Product containing cetirizine (medicinal product) |
Medicinal product Only (MPO)
775140005 | Product containing only cetirizine (medicinal product) |
Medicinal product precise only (MPP)
(not represented) | Product containing only cetirizine hydrochloride (medicinal product) |
Medicinal product Form (MPF)
768065006 | Product containing cetirizine in oral dose form (medicinal product form) |
Medicinal product form only (MPFO)
778701007 | Product containing only cetirizine in oral dose form (medicinal product form) |
Clinical drug
320818006 | Product containing precisely cetirizine hydrochloride 10 milligram/1 each conventional release oral tablet (clinical drug) |
IS-A
MP/MPO:
• Active ingredient (substance)
MPP:
• Precise active ingredient (substance)
MPF/MPFO:
• Active ingredient (substance)
• Dose form (pharmaceutical dose form)
CD:
• Precise Active ingredient (substance)
• Dose form (pharmaceutical dose form)
• Basis of strength (substance)
• Strength units (unit of measure)
• Strength values (numbers)
• Units of presentation (units of presentation)

Figure 1– SNOMED CT model for the representation of medicinal products showing the six types of entities defined in the model, along with their definitional features and examples from the SNOMED CT terminology

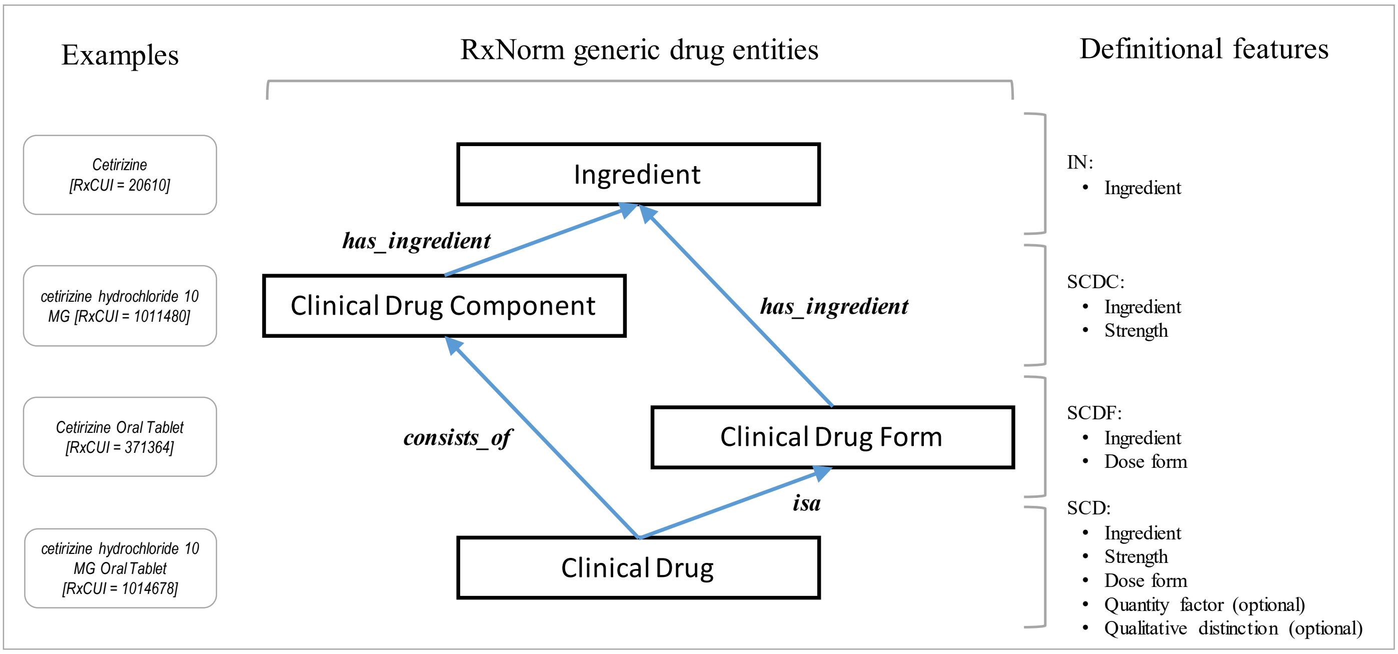

Examples
RxNorm generic drug entities
Definitional features
Cetirizine [RxCUI = 20610]
cetirizine hydrochloride 10 MG [RxCUI = 1011480]
Cetirizine Oral Tablet [RxCUI = 371364]
cetirizine hydrochloride 10 MG Oral Tablet [RxCUI = 1014678]
Ingredient
has_ingredient
Clinical Drug Component
has_ingredient
consists_of
Clinical Drug Form
isa
Clinical Drug
IN:
• Ingredient
SCDC:
• Ingredient
• Strength
SCDF:
• Ingredient
• Dose form
SCD:
• Ingredient
• Strength
• Dose form
• Quantity factor (optional)
• Qualitative distinction (optional)

Figure 2– Simplified RxNorm model for the representation of generic medicinal products showing the four types of entities defined in the model, along with their definitional features and examples from the RxNorm terminology

Closed-worldview are "closed" with respect to their active ingredient(s). More specifically, medicinal product and medicinal product form entities are closed with respect to their active ingredient(s), while medicinal product precisely and clinical drug entities are closed with respect to their precise active ingredient(s).

There are no hierarchical relations among substances. However, there is a "modification of" relation between a modified substance (e.g., ester or salt) and the corresponding base substance (e.g., between Atorvastatin calcium and Atorvastatin). Modified substances can be further modified.

IDMP requires that dose forms be defined in reference to a list of dose forms from the European Directorate for Quality in Medicines (EDQM). EDQM distinguishes between dose forms and units of presentation. Units of presentation are used to express the strength and quantity in countable entities, while dose forms correspond to the physical structure of the medicinal product.

In accordance with requirements from IDMP, strength units in SNOMED CT are aligned with the international standard for units of measure, UCUM (Unified Code for Units of Measure).

Finally, depending on the unit of presentation, strength can be represented as concentration strength, presentation strength or both.

The RxNorm model

Created in 1992, RxNorm is a normalized terminology for clinical drugs in the U.S. RxNorm represents both generic drugs and branded drugs, as well as packs (11). The full model of RxNorm contains ten entities, five for generic drug entities and five for branded drugs entities. For comparison with SNOMED CT, we only present RxNorm generic drug entities and also omit packs.

The simplified RxNorm model for generic drug entities includes four entities (Figure 2):

- ***Ingredient***, including base ingredient (IN), precise ingredient (PIN), and multi-ingredient (MIN) (e.g., IN: *Cetirizine [RxCUI = 20610]*, PIN: *cetirizine hydrochloride [RxCUI = 203150]*, MIN: *Cetirizine / Pseudoephedrine [RxCUI = 352367]*)
- ***Clinical drugs component*** (SCDC), combining ingredient and strength (e.g., *cetirizine hydrochloride 10 MG [RxCUI = 1011480]*)
- ***Clinical drugs form*** (SCDF), combining ingredient and dose form (e.g., *Cetirizine Oral Tablet [RxCUI = 371364]*)
- ***Clinical drug*** (SCD), combining ingredient, strength and dose form (e.g., *cetirizine hydrochloride 10 MG Oral Tablet [RxCUI = 1014678]*)

The representation of these entities relies on three mandatory and two optional definitional features:

- Mandatory definitional features:
  - ingredient (IN/PIN/MIN) (e.g., IN: *Cetirizine [RxCUI = 20610]*, PIN: *cetirizine hydrochloride [RxCUI = 203150]*, MIN: *Cetirizine / Pseudoephedrine [RxCUI = 352367]*)
  - dose form (DF) (e.g., *Oral Tablet [RxCUI = 317541]*)
  - strength (e.g., *10 MG*)
- Optional definitional features (see below for examples):
  - quantity factor (QF)
  - qualitative distinction (QD)

Strength in RxNorm is normalized. In its units of measure (e.g., for volume, weight, surface), RxNorm uses one unit for each type quantity (e.g., milligram for weight rather than gram or microgram).

The representation of dose forms in RxNorm is not based on a specific standard (12). It is also important to note that the SCDs

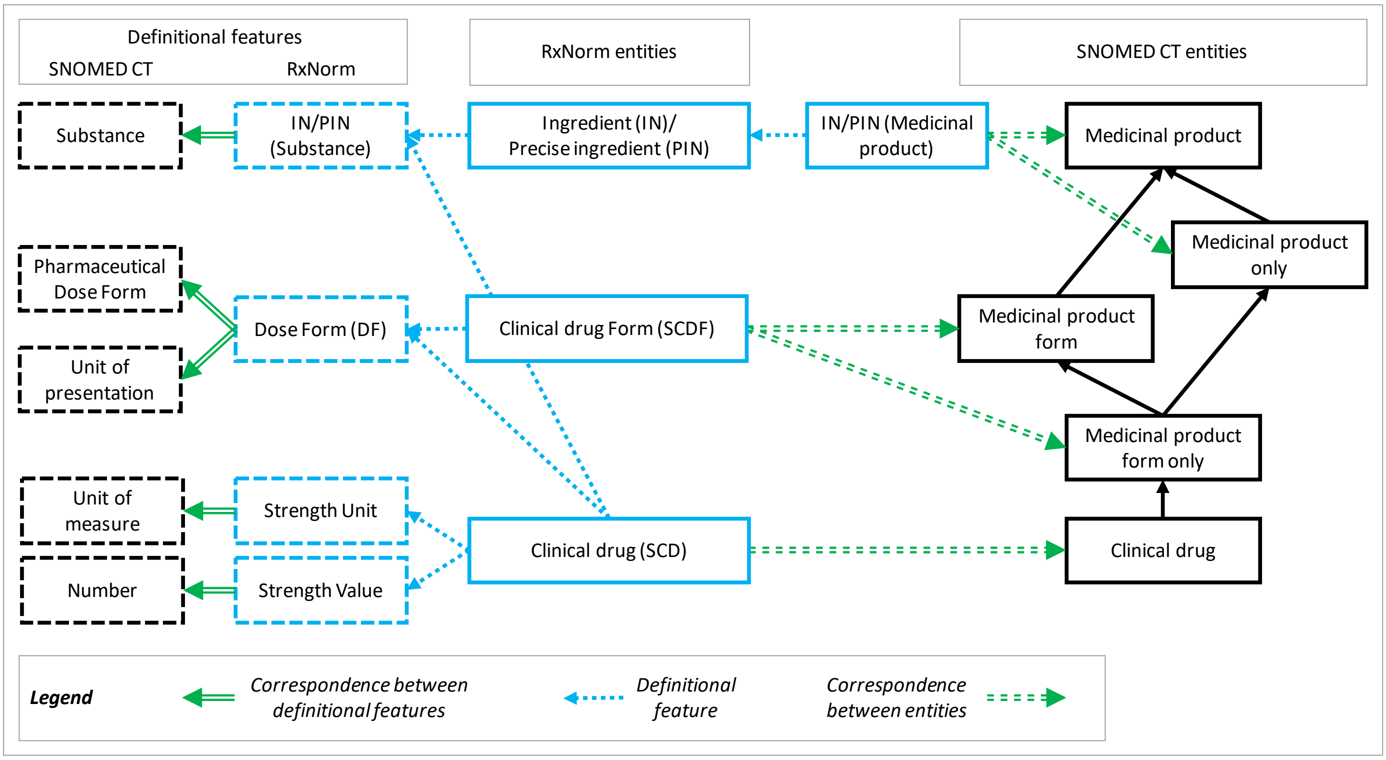


Figure 3– Correspondence between the RxNorm and SNOMED CT models

and SCDCs refer to the basis of strength substance (e.g., cetirizine hydrochloride), while SCDFs refer to the base ingredient (e.g., cetirizine). Of note, ingredients in RxNorm can (purposely) be understood as either the substance contained in a medicinal product as active ingredient (e.g., "cetirizine the substance") or the class of all medicinal products containing this substance as active ingredient. Precise ingredients (PINs) generally correspond to modified forms of the corresponding base ingredients (INs). PINs cannot be further modified.

In addition, RxNorm does not explicitly have a notion of "worldview" (i.e., open or closed worldview) for its entities. While clinical drugs implicitly refer to a closed worldview, ingredients, clinical drug components and clinical drug forms can be understood in both open and closed worldview, leaving it to queries to distinguish between the two.

Finally, the Quantity Factor (QF) is a number followed by a unit of measure corresponding to vial sizes or patch durations (e.g., "12H"). RxNorm does not explicitly state whether strength is expressed as presentation strength or concentration strength. Presentation strength can be derived from concentration strength by multiplying the concentration strength by the quantity factor. (For example, if the concentration strength is 1MG/ML and the QF is 2ML, the presentation strength is 2MG/2ML). The Qualitative Distinction (QD) corresponds to some qualitative characteristic of a drug outside the main definitional features (e.g., "sugar free" and "abuse-deterrent"). QD and QF are optional modifiers used in RxNorm to define medicinal products when it is clinically relevant to identify such distinctions (12).

Comparison of the RxNorm and SNOMED CT models

To compare the two models, we manually establish equivalences between their entities and between their definitional features, based on our analysis of the two models.

First, we need to disambiguate the notion of ingredient in RxNorm (IN,PIN, MIN), because, as mentioned earlier, it can be understood as either a substance or a class of medicinal products. Therefore, as shown in Figure 3, ingredients in RxNorm correspond to SNOMED CT medicinal products (in open and closed worldview) or to SNOMED CT substances, which are active ingredients of SNOMED CT medicinal products. In practice, RxNorm ingredients are often associated with multiple SNOMED CT entities, typically with one substance entity and one medicinal product entity. Disambiguation consists in identifying which SNOMED CT entity comes from the substance hierarchy (and treating it as a value for the definitional feature "active ingredient"), while the SNOMED CT entity corresponding to an entity from the medicinal product hierarchy is marked as an asserted equivalence for the RxNorm medicinal product entity.

RxNorm does not formally have the notion of "unit of presentation". Units of presentation are implicitly represented through dose forms in RxNorm, whereas the two notions are represented separately in SNOMED CT. For example, in SNOMED CT, tablet is the logical "unit of presentation" of the conventional release oral tablet, while the two are conflated in the RxNorm dose form "Oral Tablet". Therefore, RxNorm dose forms generally correspond to pairs of a pharmaceutical dose form and a unit of presentation in SNOMED CT.

In addition, there are no materialized entities for SCDCs in SNOMED CT. Instead, strength and basis of strength substance are associated as part of the definition of a clinical drug in

SNOMED CT. Therefore, SCDCs cannot be related to entities in SNOMED CT, but their defining features are represented as part of clinical drug entities.

SCDs in RxNorm are equivalent to clinical drugs in SNOMED CT as they essentially share the same definitional features. The quantity factor in RxNorm has no direct equivalent in SNOMED CT, but QF information is implicitly represented in the presentation strength. In contrast, qualitative distinctions are absent from the SNOMED CT model.

While RxNorm only represents one level of modification (between PIN and IN), SNOMED CT can represent arbitrary levels of modification among substances.

Both RxNorm and SNOMED CT have the notion of concentration strength and presentation strength. However, RxNorm emphasizes concentration strength (from which presentation strength can be calculated using the quantity factor), whereas SNOMED CT explicitly represent both presentation strength and concentration strength when necessary.

Finally, RxNorm normalizes all quantities to one unit (per type of quantity), whereas SNOMED CT uses units that are most clinically appropriate (following IDMP requirements). For example, RxNorm uses 0.001 milligram and SNOMED CT 1 microgram. This difference merely reflects differences in editorial guidelines, as conversion between the two is trivial.

Discussion

Findings. Not surprisingly, the models used by RxNorm and SNOMED CT for representing medicinal products are fairly similar and essentially compatible. Both models share major definitional features including ingredient (or substance), strength and dose form. Only the qualitative distinction feature of RxNorm has no correspondence at all in SNOMED CT.

SNOMED CT is more rigorous and better aligned with international standards. In SNOMED CT, differences tend to be made explicit, e.g., between a substance and the class of medicinal products containing this substance as an ingredient, or between the class of all medicinal products containing only a given active ingredient and the class of all medicinal products containing at least this active ingredient . SNOMED CT also offers more flexibility with relations among substances, as opposed to a fixed precise ingredient to base ingredient relationship in RxNorm. This precision comes at the price of a more complex model, and possibly a steeper learning curve. In contrast, RxNorm contains implicit knowledge, simplifications and ambiguities, but its model is simpler.

With features, such as explicit closed worldview for clinical drug entities, use of standard dose forms from EDQM, use of UCUM units, and use of clinically appropriate strength values, SNOMED CT shows better compliance with international standards (namely IDMP) than RxNorm does.

Consequences on alignment. Since their models are largely compatible, medicinal products from RxNorm and SNOMED CT are expected to be interoperable. However, specific aspects of the alignment between the two models require particular attention.

The values of **ingredient** can be aligned rather trivially (after disambiguation between the two meanings of RxNorm ingredients, substance and class of medicinal products containing this substance as an ingredient).

Strength entities require minimal attention, specifically for converting RxNorm “fixed unit” into the clinically appropriate unit used in SNOMED CT. Simple arithmetic is also required to convert concentration strength and quantity factor in RxNorm to presentation strength in SNOMED CT wherever appropriate.

In contrast, aligning **dose forms** requires more analysis, as RxNorm dose forms generally correspond to pairs of a pharmaceutical dose form and a unit of presentation in SNOMED CT.

The absence of correspondence for **qualitative distinction** in SNOMED CT may lead to multiple clinical drugs in RxNorm mapping to a single clinical drug in SNOMED CT. For example, the distinction between *Cholestyramine Resin 4000 MG Powder for Oral Suspension [RxCUI = 848943]* and its sugar-free form *Sugar-Free Cholestyramine Resin 4000 MG Powder for Oral Suspension [RxCUI = 1801279]* in RxNorm is lost in SNOMED CT. This issue is unlikely to result in clinically significant alignment errors.

The absence of materialization of the clinical drug component (SCDC) entity in SNOMED CT does not create an alignment issue, because SCDCs are essentially navigational entities in RxNorm. They are not crucial to any of the main use cases for RxNorm or SNOMED CT.

Future work. In future work, we plan to translate RxNorm into the SNOMED CT model for medicinal products. The resulting alignment would make RxNorm entities directly compatible with SNOMED CT’s. One benefit of this alignment would be to assess interoperability between RxNorm and SNOMED CT, potentially enriching SNOMED CT with clinical drugs currently specific to RxNorm. Additionally, this alignment would offer an opportunity for quality assurance by identifying cases where alignment is expected, but cannot be inferred (e.g., because of a difference in basis of strength substance for a given clinical drug between RxNorm and SNOMED CT).

Conclusion

In this investigation, we examined the similarities and differences between the representation of medicinal products in RxNorm and SNOMED CT. We established that both models share major definitional features including ingredient (or substance), strength and dose form. Because of subtle differences between the two models, specific aspects of their alignment require particular attention.

Acknowledgment

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. The authors would like to thank the developers of RxNorm and the contributors to the SNOMED International Drug Model Working Group for useful discussion We are particularly grateful to Tammy Powell and Chris Hui (RxNorm), and Julie James, Jim Case, Yongsheng Gao, Emma Melhuish, Toni Morrison, Guillermo Reynoso, Farzaneh Ashrafi and Phuong Skovgaard (SNOMED CT). We also thank NLM colleagues, Lee Peters, Robert Wynne and Phill Wolf, for useful discussion.

Address for correspondence

Jean.nikiema@u-bordeaux.fr
Olivier.bodenreider@nih.gov

References

1. Lupse O-S, Chirila C-B, Stoicu-tivadar L. Harnessing Ontologies to Improve Prescription in Pediatric Medicine. Studies in Health Technology and Informatics. 2018;97–101.
2. Farrish S, Grando A. Ontological approach to reduce complexity in polypharmacy. AMIA Annu Symp Proc. 2013;2013:398–407.
3. Lai EC-C, Ryan P, Zhang Y, Schuemie M, Hardy NC, Kamijima Y, et al. Applying a common data model to Asian databases for multinational pharmacoepidemiologic studies: opportunities and challenges. Clinical Epidemiology. 2018 Jul;Volume 10:875–85.
4. European Medicines Agency. Introduction to ISO Identification of Medicinal Products, SPOR programme [Internet]. 2016. Available from: https://www.ema.europa.eu/documents/other/introduction-iso-identification-medicinal-products-spor-programme_en.pdf
5. Dhavle AA, Ward-Charlerie S, Rupp MT, Kilbourne J, Amin VP, Ruiz J. Evaluating the implementation of RxNorm in ambulatory electronic prescriptions. Journal of the American Medical Informatics Association. 2016 Apr;23(e1):e99–107.
6. Liu S, Wei Ma, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Professional. 2005 Sep;7(5):17–23.
7. Wang L, Zhang Y, Jiang M, Wang J, Dong J, Liu Y, et al. Toward a normalized clinical drug knowledge base in China—applying the RxNorm model to Chinese clinical drugs. Journal of the American Medical Informatics Association. 2018 Jul 1;25(7):809–18.
8. Hanna J, Joseph E, Brochhausen M, Hogan WR. Building a drug ontology based on RxNorm and other sources. Journal of Biomedical Semantics. 2013;4(1):44.
9. Bodenreider O, James J. The New SNOMED CT International Medicinal Product Model. In: Proceedings of the International Conference on Biological Ontology (ICBO 2018). Oregon, USA,; 2018.
10. Héja G, Surján G, Varga P. Ontological analysis of SNOMED CT. BMC Medical Informatics and Decision Making. 2008;8(Suppl 1):S8.
11. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. Journal of the American Medical Informatics Association. 2011 Jul;18(4):441–8.
12. Bodenreider O, Cornet R, Vreeman D. Recent Developments in Clinical Terminologies — SNOMED CT, LOINC, and RxNorm. Yearbook of Medical Informatics. 2018 Aug;27(01):129–39.

Appendix D

Templates for RxNorm translation

D.1 template of medicinal product in open world

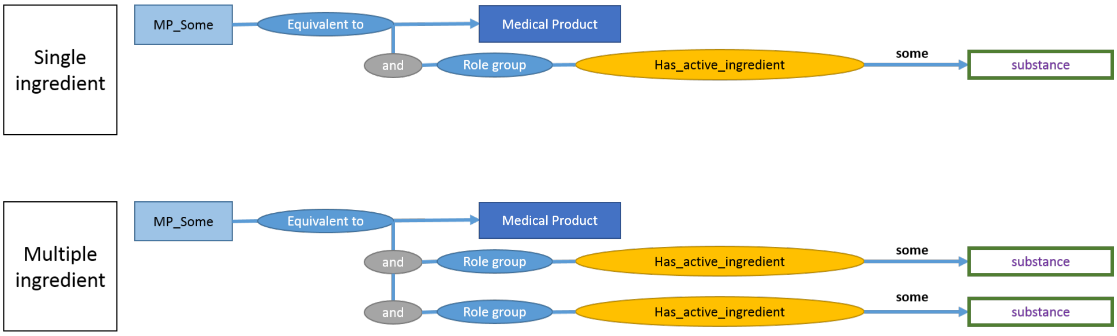
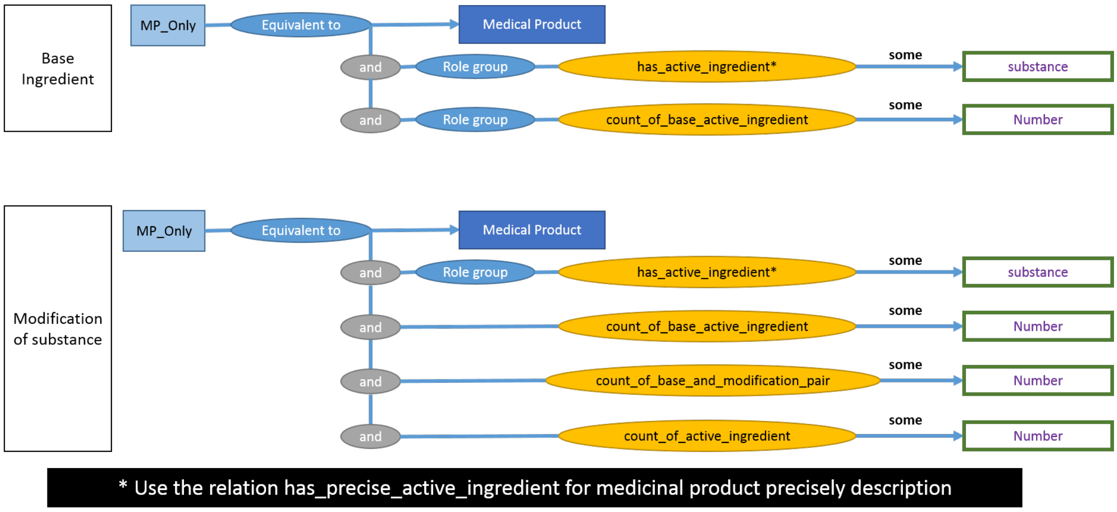


Figure D.1 – template of medicinal product in open world view for medicinal product with single and multiple ingredients

D.2 template of medicinal product in closed world: single ingredient



* Use the relation has_precise_active_ingredient for medicinal product precisely description

Figure D.2 – template of medicinal product in closed world view for medicinal product with single ingredient

D.3 template of medicinal product in closed world: multiple ingredients

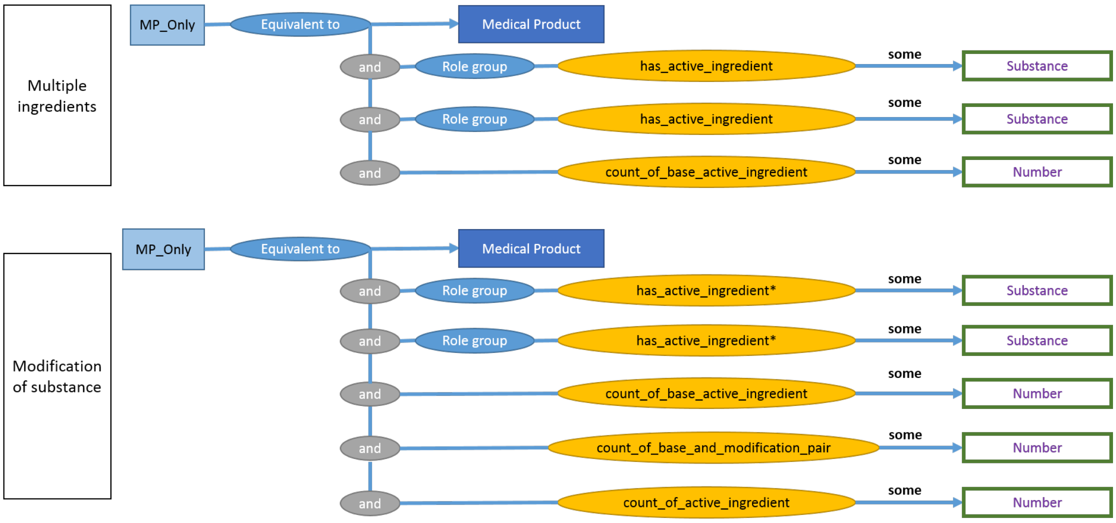


Figure D.3 – template of medicinal product in closed world view for medicinal product with multiple ingredients

D.4 template of medicinal product form

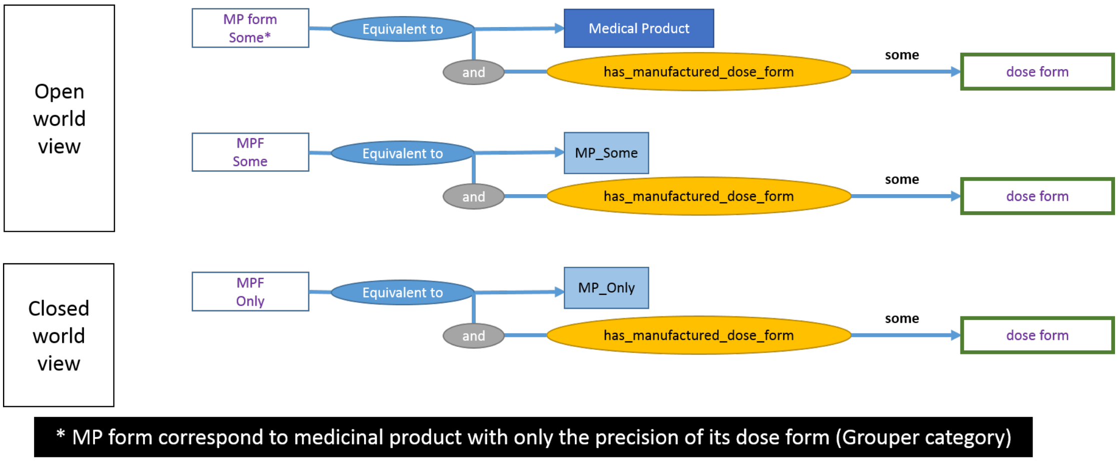
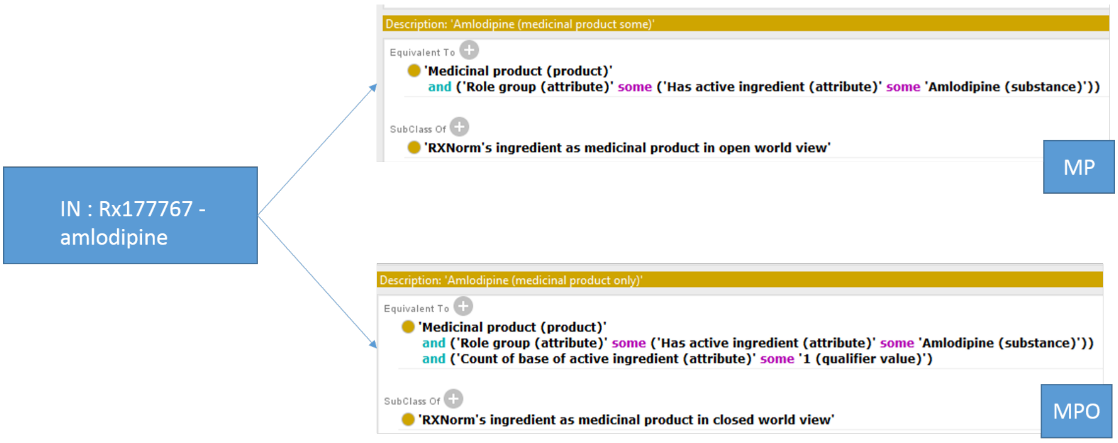

Open world view
MP form Some*
Equivalent to
Medical Product
and
has_manufactured_dose_form
some
dose form
MPF Some
Equivalent to
MP_Some
and
has_manufactured_dose_form
some
dose form
Closed world view
MPF Only
Equivalent to
MP_Only
and
has_manufactured_dose_form
some
dose form
* MP form correspond to medicinal product with only the precision of its dose form (Grouper category)

Figure D.4 – template of medicinal product form

Appendix E

Translation examples of RxNorm concepts

E.1 Instantiated ingredient IN



Description: 'Amlodipine (medicinal product some)'

Equivalent To

● 'Medicinal product (product)'
and ('Role group (attribute)' some ('Has active ingredient (attribute)' some 'Amlodipine (substance)'))

SubClass Of

● 'RXNorm's ingredient as medicinal product in open world view'

Description: 'Amlodipine (medicinal product only)'

Equivalent To

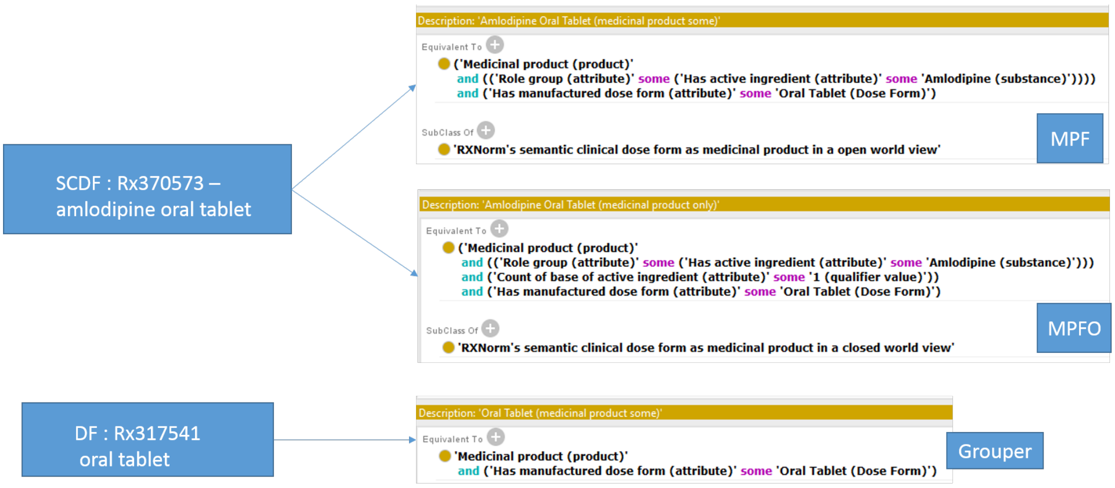
● 'Medicinal product (product)'
and ('Role group (attribute)' some ('Has active ingredient (attribute)' some 'Amlodipine (substance)'))
and ('Count of base of active ingredient (attribute)' some '1 (qualifier value)')

SubClass Of

● 'RXNorm's ingredient as medicinal product in closed world view'

Figure E.1 – Example of IN translation

E.2 Instantiated DF and SCDF



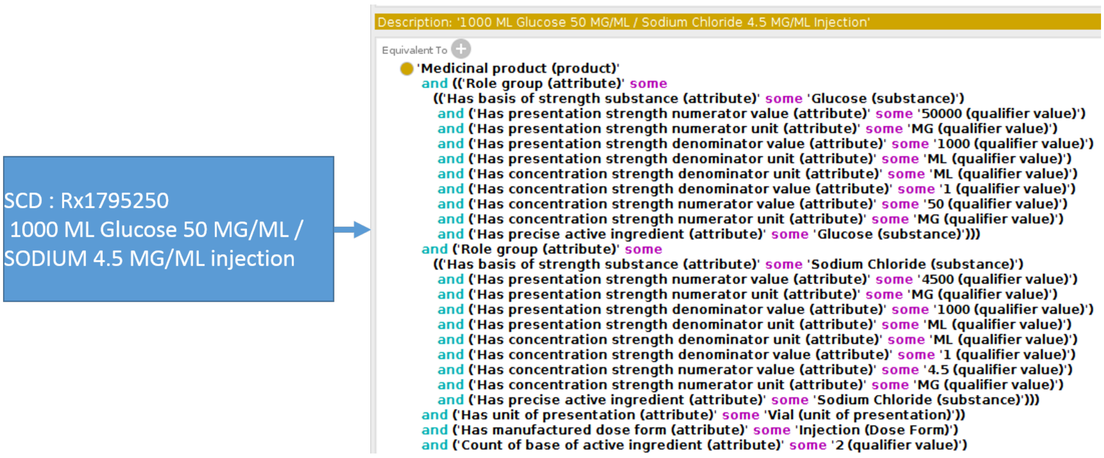
```
Description: 'Amlodipine Oral Tablet (medicinal product some)'
Equivalent To
 ('Medicinal product (product)'
   and (('Role group (attribute)' some ('Has active ingredient (attribute)' some 'Amlodipine (substance)'))))
   and ('Has manufactured dose form (attribute)' some 'Oral Tablet (Dose Form)')
SubClass Of
 'RXNorm's semantic clinical dose form as medicinal product in a open world view'
```

```
Description: 'Amlodipine Oral Tablet (medicinal product only)'
Equivalent To
 ('Medicinal product (product)'
   and (('Role group (attribute)' some ('Has active ingredient (attribute)' some 'Amlodipine (substance)')))
   and ('Count of base of active ingredient (attribute)' some '1 (qualifier value)'))
   and ('Has manufactured dose form (attribute)' some 'Oral Tablet (Dose Form)')
SubClass Of
 'RXNorm's semantic clinical dose form as medicinal product in a closed world view'
```

```
Description: 'Oral Tablet (medicinal product some)'
Equivalent To
 'Medicinal product (product)'
   and ('Has manufactured dose form (attribute)' some 'Oral Tablet (Dose Form)')
```

Figure E.2 – Example of DF and SCDF translation

E.3 Instantiated SCD



```
Description: '1000 ML Glucose 50 MG/ML / Sodium Chloride 4.5 MG/ML Injection'
Equivalent To
 'Medicinal product (product)'
   and (('Role group (attribute)' some
     (('Has basis of strength substance (attribute)' some 'Glucose (substance)')
      and ('Has presentation strength numerator value (attribute)' some '50000 (qualifier value)')
      and ('Has presentation strength numerator unit (attribute)' some 'MG (qualifier value)')
      and ('Has presentation strength denominator value (attribute)' some '1000 (qualifier value)')
      and ('Has presentation strength denominator unit (attribute)' some 'ML (qualifier value)')
      and ('Has concentration strength denominator unit (attribute)' some 'ML (qualifier value)')
      and ('Has concentration strength denominator value (attribute)' some '1 (qualifier value)')
      and ('Has concentration strength numerator value (attribute)' some '50 (qualifier value)')
      and ('Has concentration strength numerator unit (attribute)' some 'MG (qualifier value)')
      and ('Has precise active ingredient (attribute)' some 'Glucose (substance)')))
   and ('Role group (attribute)' some
     (('Has basis of strength substance (attribute)' some 'Sodium Chloride (substance)')
      and ('Has presentation strength numerator value (attribute)' some '4500 (qualifier value)')
      and ('Has presentation strength numerator unit (attribute)' some 'MG (qualifier value)')
      and ('Has presentation strength denominator value (attribute)' some '1000 (qualifier value)')
      and ('Has presentation strength denominator unit (attribute)' some 'ML (qualifier value)')
      and ('Has concentration strength denominator unit (attribute)' some 'ML (qualifier value)')
      and ('Has concentration strength denominator value (attribute)' some '1 (qualifier value)')
      and ('Has concentration strength numerator value (attribute)' some '4.5 (qualifier value)')
      and ('Has concentration strength numerator unit (attribute)' some 'MG (qualifier value)')
      and ('Has precise active ingredient (attribute)' some 'Sodium Chloride (substance)')))
   and ('Has unit of presentation (attribute)' some 'Vial (unit of presentation)'))
   and ('Has manufactured dose form (attribute)' some 'Injection (Dose Form)')
   and ('Count of base of active ingredient (attribute)' some '2 (qualifier value)')
```

Figure E.3 – template of medicinal product in closed world view for medicinal product with multiple ingredients

Contents

Integrating heterogeneous biomedical knowledge through a model based on support ontologies

Abstract

Dans le domaine de la santé, il existe un nombre très important de sources de connaissances, qui vont de simples terminologies, classifications et vocabulaires contrôlés à des représentations très formelles, que sont les ontologies. Cette hétérogénéité des sources de connaissances pose le problème de l'utilisation secondaire des données, et en particulier de l'exploitation de données hétérogènes dans le cadre de la médecine personnalisée ou translationnelle. En effet, les données à utiliser peuvent être codées par des sources de connaissances décrivant la même notion clinique de manière différente ou décrivant des notions distinctes mais complémentaires. Pour répondre au besoin d'utilisation conjointe des sources de connaissances encodant les données de santé, nous avons étudié trois processus permettant de répondre aux conflits sémantiques (difficultés résultant de leur mise en relation) : (1) l'alignement qui consiste à créer des relations de mappings (équivalence et/ou subsumption) entre les entités des sources de connaissances, (2) l'intégration qui consiste à créer des mappings et à organiser les autres entités dans une même structure commune cohérente et, enfin, (3) l'enrichissement sémantique de l'intégration qui consiste à créer des mappings grâce à des relations transversales en plus de celles d'équivalence et de subsumption. Dans un premier travail, nous avons aligné la terminologie d'interface du laboratoire d'analyses du CHU de Bordeaux à la LOINC. Deux étapes principales ont été mises en place : (i) le prétraitement des libellés de la terminologie locale qui comportaient des troncatures et des abréviations, ce qui a permis de réduire les risques de survenue de conflits de nomenclature, (ii) le filtrage basé sur la structure de la LOINC afin de résoudre les différents conflits de confusion. Deuxièmement, nous avons intégré RxNorm à la sous-partie de la SNOMED CT décrivant les connaissances sur les médicaments afin d'alimenter la SNOMED CT avec les entités de RxNorm. Ainsi, les médicaments dans RxNorm ont été décrits en OWL grâce à leurs éléments définitionnels (substance, unité de mesure, dose, etc.). Nous avons ensuite fusionné cette représentation de RxNorm à la structure de la SNOMED CT, résultant en une nouvelle source de connaissances. Nous avons ensuite comparé les équivalences inférées (entre les entités de RxNorm et celles de la SNOMED CT) grâce à cette nouvelle structure avec les équivalences préétablies de manière morphosyntaxique par RxNorm. Notre méthode a résolu des conflits de nomenclature mais était confrontée à certains conflits de confusion et d'échelle permettant ainsi de mettre en évidence des éléments d'amélioration dans RxNorm et la SNOMED CT. Finalement, nous avons réalisé une intégration sémantiquement enrichie de la CIM10 et de la CIMO3 en utilisant la SNOMED CT comme support. La CIM10 décrivant des diagnostics et la CIMO3 décrivant cette notion suivant deux axes différents (celui des lésions histologiques et celui des localisations anatomiques), nous avons utilisé la structure de la SNOMED CT pour retrouver des relations transversales entre les concepts de la CIM10 et de la CIMO3 (résolution de conflits ouverts). Au cours du processus, la structure de la SNOMED CT a également été utilisée pour supprimer les mappings erronés (conflits de nomenclature et de confusion) et désambiguïser les cas de mappings multiples (conflits d'échelle).

Keywords: semantic integration, biomedical terminology, support ontology

Bordeaux Population Health – Research Center (BPH)
INSERM U1219, ERIAS – Université de Bordeaux , Case 11 – 146 rue Léo Saignat – 33076 Bordeaux cedex – France

Integrating heterogeneous biomedical knowledge through a model based on support ontologies

Abstract

In the biomedical domain, there are almost as many knowledge resources in health as there are application fields. These knowledge resources, described according to different representation models and for different contexts of use, raise the problem of complexity of their interoperability, especially for actual public health problematics such as personalized medicine, translational medicine and the secondary use of medical data. Indeed, these knowledge resources may represent the same notion in different ways or represent different but complementary notions. For being able to use knowledge resources jointly, we studied three processes, which can overcome semantic conflicts (difficulties encountered when relating distinct knowledge resources): the alignment, the integration and the semantic enrichment of the integration. The alignment consists in creating a set of equivalence or subsumption mappings between entities from knowledge resources. The integration aims not only to find mappings but also to organize all knowledge resource entities into a unique and coherent structure. Finally, the semantic enrichment of integration consists in finding all the required mapping relations between entities of distinct knowledge resources (equivalence, subsumption, transversal and, failing that, disjunction relations). In this frame, we firstly realized the alignment of laboratory tests terminologies: LOINC and the local terminology of Bordeaux hospital. We pre-processed the noisy labels of the local terminology to reduce the risk of naming conflicts. Then, we suppressed erroneous mappings (confounding conflicts) using the structure of LOINC. Secondly, we integrated RxNorm to SNOMED CT. We constructed formal definitions for each entity in RxNorm by using their definitional features (active ingredient, strength, dose form, etc.) according to the design patterns proposed by SNOMED CT. We then integrated the constructed definitions into SNOMED CT. The obtained structured was classified and the inferred equivalences between RxNorm and SNOMED CT were compared to morphosyntactic mappings. Our process resolved some cases of naming conflicts but was confronted to confounding or scaling conflicts highlighting the needs of improvement in RxNorm and SNOMED CT. Finally, we performed a semantically enriched integration of ICD-10 and ICD-O3 using SNOMED CT as support. As ICD-10 describes diagnoses and ICD-O3 describes this notion according to two different axes (i.e., histological lesions and anatomical structures), we used the SNOMED CT structure to identify transversal relations between their entities (resolution of open conflicts). During the process, the structure of the SNOMED CT was also used to suppress erroneous mappings (naming and confusion conflicts) and disambiguate multiple mappings (scale conflicts).

Keywords: semantic integration, biomedical terminology, support ontology