Enjeux liés aux ressources termino-ontologiques en santé

Fleur Mougin

université de BORDEAUX

Enjeux liés aux ressources termino-ontologiques en santé

Habilitation à diriger des recherches

présentée et soutenue publiquement le 19 juin 2019 par

Fleur Mougin

Composition du jury

Nathalie Aussenac-Gilles, Directrice de recherche CNRS, IRIT, Toulouse (rapportrice)
Guillaume Blin, Professeur, Université de Bordeaux (examinateur)
Anita Burgun, Professeur - Praticien hospitalier, Université Paris Descartes (examinatrice)
Geneviève Chêne, Professeur - Praticien hospitalier, Université de Bordeaux (examinatrice)
Olivier Dameron, Maître de conférences - HDR, Université de Rennes 1 (rapporteur)
Pierre Zweigenbaum, Directeur de recherche CNRS, LIMSI, Orsay (rapporteur)

Remerciements

Je tiens tout d'abord à remercier vivement Nathalie Aussenac-Gilles, Olivier Dameron et Pierre Zweigenbaum d'avoir accepté de rapporter mon travail malgré leur emploi du temps très chargé. Je suis très heureuse de bénéficier de leur avis sur mes travaux de recherche et j'aimerais que cela soit l'occasion de mettre en place des collaborations futures.

Je remercie Anita Burgun de participer à mon jury et je lui suis infiniment reconnaissante de m'avoir initiée à la recherche et de m'avoir accordée sa confiance dès le début de ma carrière.

Je me réjouis de compter Geneviève Chêne parmi les membres de mon jury. Sa manière de diriger l'ISPED et sa vision de la santé publique sont une source d'inspiration pour moi.

Je remercie Guillaume Blin d'avoir été aussi réactif et enthousiaste à l'idée d'examiner mon travail.

Merci à Jean-François Dartigues pour son aide au projet SemBiP.

Merci à Roger Salamon d'avoir cru en nous pour la direction de la thèse de Khadim et pour son perpétuel soutien à l'équipe ERIAS.

Je remercie Rodolphe Thiébaut pour les collaborations que nous avons menées ensemble et pour son support au recrutement d'Aarón.

Merci à Olivier Bodenreider qui m'a tant appris sur les rouages du métier de chercheur.

Je veux exprimer ma joie d'avoir rencontré Natalia Grabar et, grâce à elle, Thierry Hamon avec qui il est si facile et agréable de travailler.

Merci à mes collègues Lina Soualmia et Sandra Bringuay pour leurs conseils de rédaction.

Je veux remercier tout particulièrement mes doctorants (actuels et ancien), Aarón, Jean Noël et Khadim, ainsi que Georgeta que j'ai pris plaisir à encadrer et sans qui ce manuscrit serait bien vide...

Merci à Gayo et à Vianney avec qui j'apprécie beaucoup d'effectuer des co-encadrements. Nos échanges sont sans cesse enrichissants et stimulants, que ce soit professionnellement que personnellement !

J'ai une pensée particulière pour Frantz avec qui je partage un bureau depuis mon arrivée à l'ISPED et qui me prête toujours une oreille attentive.

Nous travaillons plus épisodiquement ensemble mais c'est avec engouement : Sébastien, Romain, Bruno. Je me réjouis de pouvoir travailler avec eux dont j'admire la fougue.

Je pense aussi à mes collègues de l'ISPED et du BPH qui sont toujours prêts à discuter de toutes sortes de sujets : Alioum, Amandine, Barbara, Boris, Carole D, Carole Q, Cécilia, Hélène, Marta, Nancy, Pierre, Robin, et l'équipe pédagogique plus généralement.

Merci aux copines de goguettes : Cécile, Élise, Gaëlle, Karen et Valérie. C'est sympa d'échanger sur nos pratiques professionnelles sans se prendre au sérieux mais tellement plus de partager avec elles des moments de détente à la cafet', dans le patio de l'université ou à l'Arrozoir...

Merci à Nana qui compte tant depuis notre rencontre à Cercedilla... et avec qui je ne désespère pas de collaborer ! Toutes ces choses que nous avons en commun nous permettent de prendre du recul par rapport à notre métier, mais elles devraient aussi nous offrir la possibilité de réaliser des travaux de recherche ensemble, ne serait-ce que pour nous voir plus souvent.

Je suis extrêmement reconnaissante à Patricia pour sa relecture attentive de mon manuscrit et ses précieux conseils pour en améliorer le contenu. Je la remercie aussi pour nos collaborations et tous les bons moments que nous avons passés et passerons ensemble.

Merci enfin à mes parents pour leur soutien constant, et à toute ma famille. Merci à mes amis, sur qui je sais que je peux compter. J'adresse un merci spécial à ma sœur, Marie Mou, d'être à mes côtés au quotidien, attentionnée et bienveillante. Merci aussi à mes garçons, Nino et Solan, qui me divertissent si bien : quand on fait les devoirs, lors de nos parties de jeux de société endiablées, au cours de nos sorties au parc ou de nos balades en ville, pendant les vacances... Bien entendu, un grand MERCI à Richard qui me supporte, me rassure, m'encourage systématiquement et est toujours là, tout simplement !

Table des matières

CHAPITRE 1

INTRODUCTION

1.1 Contexte

Plus de 60 ans après l'introduction du terme "intelligence artificielle" par John McCarthy pour caractériser la capacité à reproduire l'intelligence humaine par une machine, on assiste depuis peu à un engouement majeur et généralisé pour cette thématique. Les stratégies nationales pour l'utilisation et le développement de l'intelligence artificielle définies par de nombreux pays au cours des deux dernières années (Figure 1.1) illustrent leur ambition d'être des acteurs importants dans ce domaine. L'explosion du nombre de publications sur ce thème dans la base de données bibliographiques MEDLINE en 2018 est un autre indicateur de l'intérêt porté à l'intelligence artificielle et témoigne de la volonté de chercheurs de spécialités autres que les mathématiques et l'informatique de s'en saisir (Figure 1.2). De nombreuses publications récentes dans le champ de la santé considèrent essentiellement les méthodes d'apprentissage (*machine learning*) quand elles font référence à l'intelligence artificielle (IA) [1], que ce soit dans le domaine de la radiologie [2], de la cardiologie [3], du vieillissement [4], et bien d'autres. Ceci est fortement lié à l'extraordinaire potentiel des données massives (*big data*) dont les caractéristiques, initialement définies comme les 3 V que sont le volume, la vélocité et la variété [5] et qui ont été largement étendues depuis [6], ont nécessité d'adapter les méthodes existantes pour pouvoir les traiter.

Cependant, les méthodes d'apprentissage sont un des pans de l'IA, au même titre que les approches basées sur la logique et les règles (parfois dénommées approches symboliques). Tandis que les méthodes statistiques d'apprentissage prennent en compte l'incertitude des données pour faire des prédictions à partir d'observations potentiellement incomplètes, les méthodes basées sur la logique se focalisent sur la complexité des données et permettent de tirer des conclusions à partir de faits. Russell a fait référence à ces deux mouvances dans [7]. Il a désigné les approches d'apprentissage sous l'appellation d'*IA moderne*, faisant écho aux nombreux travaux de recherche réalisés ces dernières années pour répondre aux enjeux du *big data*. À l'inverse, il a qualifié les approches logiques d'*IA classique*, telle que caractérisée par McCarthy. À l'origine, l'IA a en effet été définie sous le prisme de la logique, avec l'objectif de fournir aux machines des capa-

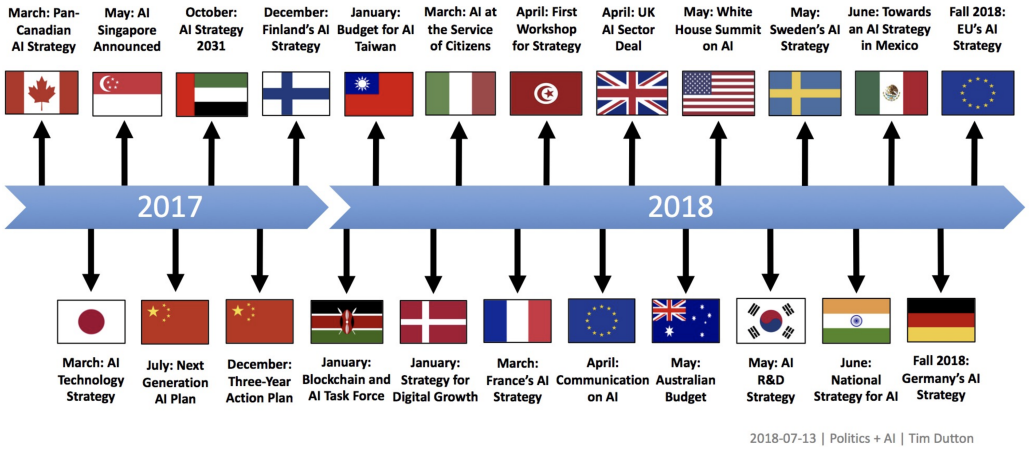

FIGURE 1.1 – Stratégies nationales mises en place pour promouvoir l'utilisation et le développement de l'intelligence artificielle entre le printemps 2017 et l'automne 2018 (source : `https://medium.com/politics-ai/an-overview-of-national-ai-strategies-2a70ec6edfd`).

cités de raisonnement comparables à celles d'un humain [8]. À l'ère de l'*IA 2.0* [9], il faut non seulement concevoir des algorithmes de prédiction à partir des nombreuses données disponibles mais aussi des méthodes permettant de structurer et d'interpréter ces données qui sont souvent non structurées [10]. Le besoin de transformer des données non structurées en connaissances relève de la représentation des connaissances. Plus largement, cette thématique de recherche vise à décrire un substitut des objets du monde réel que l'on cherche à représenter, et ce au moyen d'un langage formel interprétable par des machines pouvant ainsi effectuer des raisonnements intelligents [11]. Par ailleurs, comme l'a souligné Aussenac-Gilles, *"à la fin des années 90, la perspective s'est élargie et a donné naissance à l'**ingénierie des connaissances** [...], définie comme la partie de l'IA qui étudie et propose des concepts, méthodes, techniques et outils permettant d'acquérir, de modéliser et de formaliser des connaissances dans les organisations dans un but d'opérationnalisation, de structuration ou de gestion au sens large"* [12]. Mes travaux de recherche se positionnent dans ce cadre et concernent plus spécifiquement la conception de systèmes de représentation des connaissances, leur interopérabilité sémantique ainsi que l'évaluation de leur qualité.

Le domaine d'application choisi est celui de la **santé** pour plusieurs raisons liées à sa complexité. Tout d'abord, la richesse du vocabulaire médical pose des problèmes d'ambiguïté, de synonymie et d'imprécision [13]. De plus, certaines notions peuvent être représentées de manière atomique ou en composant plusieurs caractéristiques. Par exemple, les tumeurs peuvent être décrites comme des diagnostics, ou bien par la combinaison de leur morphologie (*e.g.*, malignes, bénignes) et de leur topographie (*e.g.*, sein, prostate). Ensuite, c'est un domaine dans lequel la production de systèmes de représentation des connaissances a toujours été naturelle, depuis Aristote en passant notamment par Galien [14]. Les premiers systèmes classificatoires de maladies sont apparus dès le 18$^{\text{ème}}$ siècle : le Genera Morborum publié en 1759 par Carl von Linné et la Nosologia Methodica publiée en 1763 par François Boissier de Sauvages de Lacroix. Aujourd'hui, on recense un nombre important de ces systèmes dans le domaine biomédical [1] [15], qui sont non seulement volumineux mais aussi de différents types car ils n'ont pas été développés pour servir un objectif unique mais pour un besoin bien précis [13, 16].

1. J'utilise les termes "domaine de la santé" et "domaine biomédical" de manière indifférenciée.

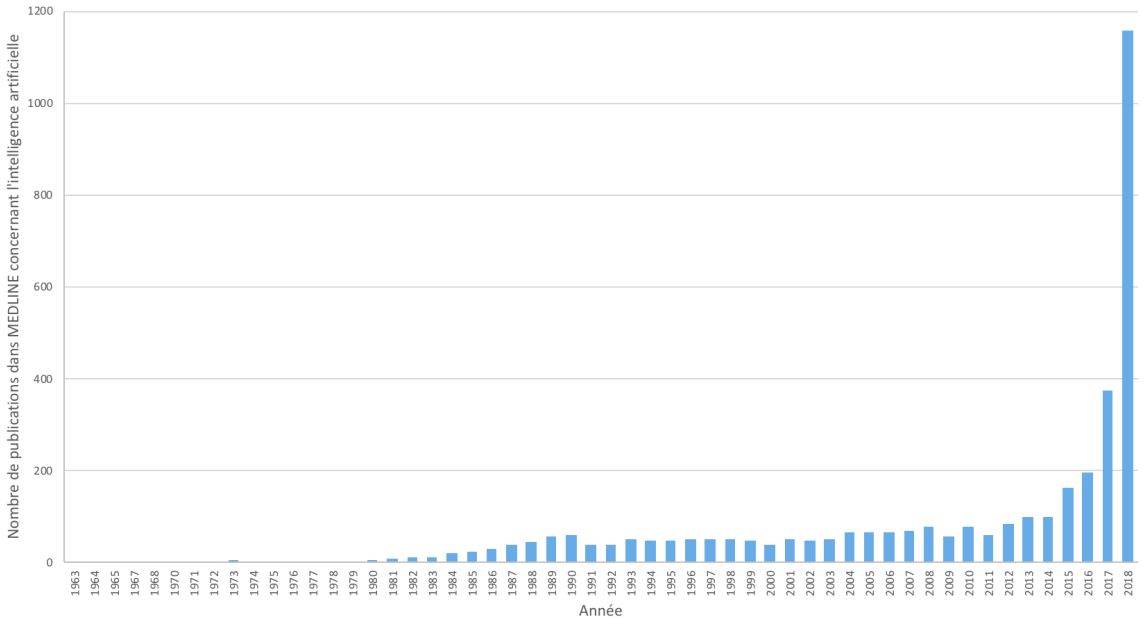

FIGURE 1.2 – Évolution du nombre de publications dans MEDLINE de 1963 à 2018 où le terme "artificial intelligence" apparaît dans le titre et/ou le résumé (*abstract*) (requête effectuée avec PubMed le 24/01/2019). Nombre total de publications de 1963 à 2018 : 3674, dont 1159 (31,5%) en 2018.

1.2 Typologies des systèmes de représentation des connaissances

De nombreuses typologies des systèmes de représentation des connaissances existent et le but n'est pas ici de donner une présentation exhaustive de celles-ci mais plutôt d'introduire les différentes appellations que l'on peut rencontrer dans le domaine de la santé. En préambule, je définis ce que j'entends par "terme" et "concept" d'après le triangle sémiotique d'Ogden et Richards au regard de l'objet du monde réel que l'on souhaite représenter [17] : un **concept** est la notion abstraite qui fait référence à cet objet tandis qu'un **terme** est un mot ou un ensemble de mots qui permettent de désigner l'objet en question.

De Keizer *et al.* ont regroupé les différents types de systèmes de représentation des connaissances utilisés en santé sous le terme générique de "systèmes terminologiques", illustrant l'importance de la composante lexicale (au travers des termes) dans ce domaine [13]. Les auteurs ont proposé la typologie suivante précisant les propriétés propres à chaque système [18] :

— une **terminologie** est le système de base dans lequel sont listés les termes d'un domaine particulier,

— un **thesaurus** est une terminologie dans laquelle les termes sont ordonnés (par ordre alphabétique, par exemple) et dans laquelle les concepts peuvent être décrits par plus d'un terme (on parle alors de **synonymes**),

— une **classification** est une structure dans laquelle les concepts sont regroupés suivant des caractéristiques communes au sein de classes. Les auteurs précisent qu'une **taxonomie** est une classification où la structure des classes est basée sur une hiérarchie de généralisation/spécialisation (dénommées par la suite **relations *is_a*** ou **relations de subsomption**),

— un **vocabulaire** ou **glossaire** est une terminologie dans laquelle les termes et/ou les concepts ont une définition textuelle,

— une **nomenclature** est une terminologie où les termes peuvent être combinés suivant des règles de composition préétablies afin de définir de nouveaux concepts,

— un **système de codage** est une terminologie dans laquelle chaque concept a un code.

Par ailleurs, les auteurs évoquent la notion d'**ontologie** en précisant que cette dénomination est parfois utilisée pour désigner indifféremment les différents types de systèmes terminologiques qui existent en santé et en fournissant la définition de Gruber, à savoir "*une spécification explicite d'une conceptualisation*" [19]. Celle-ci a été enrichie par Studer pour rendre compte de l'interopérabilité sémantique : "*une spécification formelle et explicite d'une conceptualisation partagée*" [20]. En d'autres termes, une ontologie vise à décrire dans un langage opérationnalisable la représentation mentale d'objets définie de manière consensuelle.

Une typologie des systèmes de représentation des connaissances en santé légèrement différente a été proposée par Cornet et Chute [21], reprise récemment dans [22]. Dans [21], les auteurs ont défini une terminologie comme un système constitué de concepts, reliés de manière hiérarchique, chacun ayant un identifiant unique et étant désigné par un ou plusieurs termes en langage naturel. La notion de classification est, quant à elle, plus précise car les auteurs indiquent que c'est un sytème terminologique visant à décrire un domaine de manière exhaustive, grâce à la présence de classes dites résiduelles, telles que "Hypothyroïdie, sans précision" et "Autres affections articulaires spécifiques, non classées ailleurs". Par ailleurs, une ontologie y est définie comme un système de représentation dans lequel les concepts sont reliés entre eux par des relations hiérarchiques et non hiérarchiques, parfois qualifiées de relations associatives [23] ou transversales [24], et décrit dans un langage formel basé sur les logiques de description [25]. Haendel *et al.* ont complété cette description en précisant que les ontologies (et c'est aussi le cas de certains systèmes de représentation des connaissances de plus bas niveau) décrivent de manière explicite des connaissances qui sont implicites dans des textes médicaux et peuvent servir à désambiguïser les termes se trouvant dans ces textes [22].

En dehors du domaine de la santé, des typologies ont également été données, comme celle de Bourigault *et al.* [26]. Les auteurs expliquent que les différents types de systèmes de représentation se distinguent d'après les besoins pour lesquels ils ont été créés et l'unité de connaissances qu'ils manipulent (*i.e.*, concept ou terme). Ils soulignent qu'il est indispensable de comprendre les spécificités de ces différents types de systèmes de représentation pour identifier leurs caractéristiques communes et proposer des outils génériques permettant de les manipuler, les exploiter, les évaluer, *etc.* Dans un but d'uniformisation, Bourigault et Aussenac-Gilles ont introduit le terme de **ressources termino-ontologiques** (RTOs) pour les désigner sous une même appellation [27]. Dans la suite de ce manuscrit, c'est ce terme que j'ai choisi d'utiliser pour qualifier les différents types de systèmes de représentation des connaissances utilisés en santé.

1.3 Intérêt des ressources termino-ontologiques en santé

De nombreux travaux de la communauté informatique ont recensé les rôles joués par les ontologies pour l'IA en général (*e.g.*, [28, 29]). Dans le domaine biomédical, c'est plus globalement l'intérêt des RTOs qui est présenté [15, 30, 31, 32]. Notons cependant que le terme "ontologie"

est souvent utilisé de manière générique par la communauté d'informatique médicale, ce qui est incorrect à mon sens puisque de multiples RTOs qui existent en santé ne disposent pas des caractéristiques propres à une ontologie. Dans [30], Bodenreider a réparti les différents rôles des RTOs biomédicales selon trois grands axes : (i) la gestion de connaissances, (ii) l'intégration de données, et (iii) l'aide à la décision et le raisonnement.

Ainsi, les RTOs **représentent les connaissances** d'un domaine ou d'une application particuliers qui sont exploitées en santé pour le codage de l'information médicale, l'indexation d'articles scientifiques (comme le MeSH [2] pour les publications de MEDLINE) ou encore la reconnaissance de termes dans des documents textuels. Cette reconnaissance est réalisée grâce à des outils de traitement automatique des langues, tels que MetaMap [33] ou le SIFR annotator pour les textes en français [34], qui exploitent le contenu des RTOs pour identifier les termes du domaine qu'elles représentent. Dans le domaine biologique, les RTOs sont fréquemment utilisées pour l'annotation de gènes ou de produits de gènes, à l'image de Gene Ontology Annotation [35] qui associe des termes de la RTO Gene Ontology [36] aux produits de gènes d'un organisme donné afin de décrire leurs rôles. Par ailleurs, les relations décrites entre les entités présentes dans les RTOs sont potentiellement utiles pour la recherche d'information. Cela correspond à l'expansion de requêtes (voir notamment [37] pour une revue des stratégies existantes) qui génère des résultats plus complets en incluant des documents qui concernent des termes proches du ou des termes recherchés initialement. De manière comparable, les RTOs facilitent la sélection de données, par exemple pour constituer des cohortes ou pour sélectionner un groupe spécifique de patients lors de la mise en place d'un essai clinique ou thérapeutique (grâce au codage de l'information précédemment mentionné) [38, 39].

Les RTOs peuvent également aider à l'**intégration de données**, en servant de standards d'échange ou de support à l'interopérabilité sémantique entre plusieurs systèmes d'information. Divers standards ont été proposés pour faciliter l'échange entre les systèmes de santé en offrant un modèle commun pour décrire les données cliniques, le dernier né étant FHIR [3]. Il en existe aussi dans le domaine biologique avec, entre autres, BioPAX qui permet de représenter les données concernant les voies biologiques de manière normalisée et par là-même de favoriser la réutilisation de ces données [40]. En ce qui concerne les approches d'intégration de type entrepôt de données ou basée sur un médiateur, les RTOs jouent un rôle différent. Dans le premier cas, les données issues de multiples sources sont stockées dans un unique entrepôt centralisé. Les RTOs fournissent alors la possibilité de représenter ces données suivant un vocabulaire commun (d'autres avantages des RTOs dans ce type de système d'intégration sont illustrés dans [41, 42]). Dans l'approche médiateur, les données restent dans les sources d'origine et les RTOs servent : (i) de socle au schéma global suivant lequel l'utilisateur effectue ses requêtes, et (ii) pour établir des liens entre les éléments du schéma global et ceux des schémas des sources (*e.g.*, [43, 44]). L'agrégation de données peut aussi être réalisée grâce aux RTOs. Par exemple, il est fréquent que les biologistes aient à interpréter des groupes de gènes ayant des comportements similaires. Il est alors nécessaire de considérer l'ensemble des annotations des différents gènes d'un même groupe, qui sont potentiellement très nombreuses et rendent leur interprétation manuelle impossible. Les RTOs peuvent permettre de déterminer si des liens existent entre ces annotations grâce à des

2. `https://www.nlm.nih.gov/mesh/meshhome.html`
3. `https://www.hl7.org/fhir/`

mesures de similarité sémantique [45, 46, 47] et, le cas échéant, l'agrégation des annotations résulte en un plus petit ensemble d'annotations, plus faciles à interpréter par un humain [48].

Enfin, les RTOs sont particulièrement profitables pour l'**aide à la décision** et pour effectuer du **raisonnement**. En effet, les RTOs apportent les connaissances qui sont nécessaires pour établir des règles menant à la meilleure décision dans un contexte donné [49]. Par exemple, un système d'aide à la prescription peut en avoir l'utilité [50]. En exploitant une RTO qui décrit les interactions entre des médicaments, le système est capable de fournir une alerte aux professionnels de santé en présence d'une co-prescription de médicaments entre lesquels une interaction existe. Si la RTO précise à quelle(s) classe(s) chaque médicament appartient, le système peut aussi indiquer qu'il ne faut pas prescrire un médicament donné parce qu'il fait partie d'une classe de médicaments auquel un patient est allergique [51]. D'autre part, lorsque les connaissances d'une RTO sont décrites dans un langage formel (qui est donc interprétable par une machine), il est possible de reproduire le mécanisme de déduction [52]. La catégorisation automatique d'entités dans des classes plus générales est réalisable (*e.g.*, classer un produit de gènes dans une famille de protéines d'après sa structure [53]), tout comme la prédiction d'événements (*e.g.*, déduire des blessures secondaires suite à une blessure par balle en fonction des parties anatomiques touchées au niveau du cœur [54]). Par ailleurs, la capacité de raisonnement offerte par les RTOs formelles leur est intrinsèquement utile. En effet, les raisonneurs (outils informatiques se basant sur la description formelle des RTOs pour effectuer des déductions) permettent non seulement de gérer l'évolution des RTOs en classant automatiquement de nouveaux concepts au sein de celles-ci mais aussi d'évaluer la cohérence des RTOs après avoir vérifié si leurs éléments y sont représentés correctement [55, 56]. L'exploitation des définitions logiques décrites dans les RTOs formelles favorisent aussi l'identification de correspondances entre des entités de RTOs différentes, y compris si elles décrivent des notions distinctes [57]. La capacité de raisonnement peut donc améliorer indirectement les fonctionnalités des applications basées sur des RTOs formelles.

1.4 Problématiques liées aux ressources termino-ontologiques

Pour donner un cadre aux travaux que j'ai réalisés sur les RTOs biomédicales, j'ai choisi de les positionner par rapport au cycle de vie d'une ontologie introduit il y a une vingtaine d'années par les chercheurs qui ont défini des méthodologies de conception d'ontologies. Un état de l'art de telles méthodes est disponible dans [58]. Parmi celles-ci, on peut citer METHONTOLOGY [59] proposée par Fernández-López *et al.* [59], Ontology Development 101 décrite par Noy et McGuinness [60] ou encore celle de Bachimont [61]. Ces méthodologies énumèrent les étapes nécessaires au développement d'une ontologie. METHONTOLOGY liste les sept étapes suivantes :

1. la spécification (parfois nommée "analyse des besoins") où l'on doit préciser le domaine, la portée, les utilisateurs et les utilisations de l'ontologie,
2. l'acquisition de connaissances visant à identifier les termes, les concepts et/ou les relations de l'ontologie et leur définition,
3. la conceptualisation où le but est de structurer les connaissances du domaine selon un modèle (conceptuel notamment),
4. l'intégration qui permet d'articuler l'ontologie par rapport à des ontologies existantes pouvant fournir une structuration des entités génériques (via une ontologie dite de "haut niveau") ou servir de base à la description des connaissances du domaine d'intérêt,

5. l'implémentation de l'ontologie en la décrivant selon un langage formel,
6. l'évaluation où l'on vérifie la cohérence de l'ontologie produite et son adéquation aux besoins initiaux,
7. la documentation garantissant la compréhension globale de l'ontologie et sa réutilisation.

Les méthodologies Ontology Development 101 et de Bachimont se focalisent essentiellement sur les trois premières étapes listées ci-dessus, présentant l'avantage de fournir des principes très précis pour faciliter les choix de représentation à faire lors de la conceptualisation. En fait, METHONTOLOGY est une méthodologie recouvrant plus globalement le cycle de vie d'une ontologie. Fernández-López *et al.* ont en effet introduit la maintenance de l'ontologie comme une étape supplémentaire intervenant en aval de la conception initiale et qui fait du cycle de vie d'une ontologie un processus itératif [62]. Quelques années plus tard, Corcho *et al.* ont affiné la description de ce processus en regroupant l'ensemble des étapes dans les trois catégories d'activités suivantes [63] :

— les activités qui relèvent de la **gestion** de l'ontologie où on trouve la planification, le contrôle et l'évaluation de la qualité (*quality assurance*),

— les activités qui visent au **développement** lui-même et déclinées suivant les activités à réaliser en amont (telles que l'étude de faisabilité), pendant (comme la spécification et l'implémentation) et en aval du développement (par exemple, la maintenance).

— les activités dites de **support** qui sont indispensables au développement, où on trouve notamment l'acquisition de connaissances, la documentation, l'alignement et l'intégration.

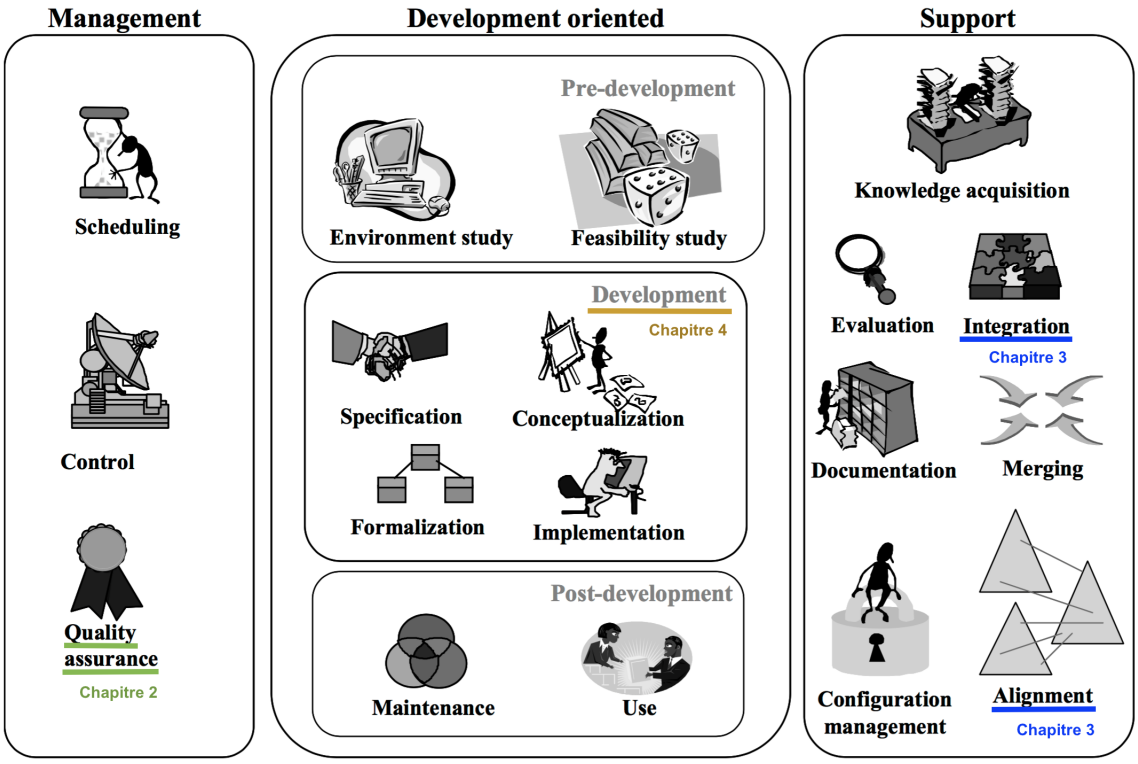

FIGURE 1.3 – Étapes du cycle de vie d'une ontologie décrites par Corcho *et al.* (source : [63]). Mes contributions portant sur l'évaluation de la qualité des RTOs, l'alignement et l'intégration ainsi que le développement de RTOs sont décrites dans les chapitres précisés sous les activités concernées.

Les contributions décrites en détails dans ce manuscrit relèvent des trois catégories d'activités décrites par Corcho *et al.* puisqu'elles concernent **l'évaluation de la qualité, le développement ainsi que l'alignement et l'intégration de RTOs en santé** (Figure 1.3). L'alignement et l'intégration sont regroupés dans le chapitre 3 sous le terme plus général d'interopérabilité sémantique englobant ces deux processus qui permettent une utilisation conjointe de RTOs au travers d'approches différentes [64].

1.5 Résumé des contributions

Dans le cadre de ma thèse intitulée "Conception d'un modèle Web sémantique appliqué à la génomique fonctionnelle" et effectuée à l'université de Rennes 1 au sein de l'EA 3888 Modélisation Conceptuelle des Connaissances Biomédicales sous l'encadrement d'Anita Burgun, j'ai développé un système d'intégration basée sur une approche médiateur [65, 66]. Celui-ci visait à offrir aux biologistes et médecins un accès centralisé et transparent à différentes sources de données biomédicales, bien que celles-ci soient distribuées et hétérogènes à de multiples niveaux. Tout d'abord, j'ai proposé une méthode d'acquisition automatique des schémas des sources de données à intégrer (dits schémas locaux) [67]. Ensuite, j'ai utilisé une RTO existante afin de définir le schéma global. Enfin, j'ai élaboré des approches automatiques pour mettre en correspondance les éléments des schémas locaux avec ceux du schéma global [68, 69, 70]. Comme le titre de ma thèse l'indiquait, la problématique d'intégration de sources de données hétérogènes relevait du Web sémantique [71, 72] tel qu'il a été défini par Tim Berners-Lee [73].

À mon arrivée à l'université de Bordeaux en septembre 2007 en tant que maître de conférences en informatique, mon rattachement recherche était dans une unité Inserm. La recherche en informatique médicale n'était pas structurée au sein d'une équipe à part entière et reposait uniquement sur un maître de conférences - praticien hospitalier et un interne de santé publique, tous deux spécialisés en informatique médicale. Nos premiers travaux communs ont débuté en 2008 grâce à trois projets européens du septième programme cadre (FP7) dont le but était de détecter les effets indésirables de médicaments à partir de données patients, certaines étant structurées et d'autres disponibles en texte libre. Dans ces projets, notre contribution a été de rendre possible la comparaison de données codées avec des RTOs différentes [74, 75, 76, 77]. Par ailleurs, j'ai continué à collaborer avec Anita Burgun et Olivier Bodenreider, qui était à l'époque chercheur à la Cognitive Science Branch du Lister Hill National Center for Biomedical Communications de la National Library of Medicine (NLM) aux États-Unis et dont il est maintenant le directeur, avec qui j'avais travaillé pendant ma thèse [78]. Nous avons mené plusieurs études sur l'alignement de RTOs et l'évaluation de leur qualité. Dans l'un de ces travaux, nous nous sommes intéressés à l'alignement de phénotypes des mammifères avec des phénotypes cliniques humains [79]. L'alignement a été établi de manière indirecte grâce à deux approches complémentaires : l'utilisation d'une RTO pivot et l'exploitation de l'annotation de gènes orthologues. Dans [80], nous avons cherché à aligner deux RTOs sur les médicaments en prenant en compte la classe pharmaceutique à laquelle ils appartenaient. Sur l'aspect évaluation, nous avons analysé la qualité des relations hiérarchiques et associatives d'une RTO liée au domaine de la cancérologie en utilisant des technologies du Web sémantique [81]. Finalement, nous avons mis en évidence l'existence de concepts polysémiques dans un vaste système intégrant plus de 150 RTOs biomédicales et déterminé la raison pour laquelle ce type de concepts existaient [82].

Avec le recrutement d'un deuxième maître de conférences en informatique en septembre 2009, Gayo Diallo, nous nous sommes structurés au sein d'une équipe émergente, l'Équipe de Recherche en Informatique Appliquée à la Santé (ERIAS), officiellement rattachée au centre de recherche Inserm Bordeaux Population Health (U1219) lors de sa création en janvier 2016. Au cours de ces années, nous avons cherché à mettre en place des collaborations locales avec des équipes de recherche en santé publique afin de faciliter notre intégration dans le centre de recherche Inserm. Avec mon collègue Gayo, nous avons obtenu un financement de 100 000€ par la Fondation Plan Alzheimer en 2011. Ce projet, en collaboration avec l'équipe Épidémiologie et Neuropsychologie du Vieillissement Cérébral du centre Inserm, visait à offrir aux professionnels de santé et aux décideurs publics un accès efficace à des articles scientifiques sur la maladie d'Alzheimer et les syndromes apparentés. Les travaux méthodologiques réalisés dans le cadre de la thèse de Khadim Dramé [83] ont d'abord consisté à concevoir une RTO bilingue français-anglais de la maladie d'Alzheimer et les syndromes apparentés : OntoAD [84, 85, 86]. Khadim a ensuite élaboré une méthode d'indexation sémantique et créé un système de recherche d'information guidée par OntoAD [87]. Dans un souci de passage à l'échelle, il a cherché à exploiter uniquement une partie des documents (*i.e.*, titre et résumé d'articles scientifiques) pour les indexer. Les concepts qui peuvent être représentatifs d'un document n'étant pas toujours explicitement mentionnés dans cette information incomplète, il a mis en œuvre une méthode de classification de documents basée sur l'algorithme des k plus proches voisins [88, 89]. Celle-ci a été évaluée avec succès sur plusieurs centaines de milliers de documents biomédicaux fournis dans le cadre du défi international CLEF "Cross-Language Evaluation for eHealth Document Analysis" en 2014 [87].

Par ailleurs, j'ai pris contact avec des chercheurs du laboratoire bordelais de recherche en informatique (LaBRI, UMR CNRS 5800) pour maintenir un lien étroit avec la communauté informatique. Depuis 2015, je suis membre associée du LaBRI et affiliée à l'équipe Bench to Knowledge and Beyond (BKB), anciennement nommée Modèles et Algorithmes pour la Bioinformatique et la Visualisation d'informations (MABioVis). J'ai ainsi mis en place des collaborations avec des informaticiens spécialistes en visualisation d'information et avec des bioinformaticiens. Ce rapprochement s'est concrétisé par deux activités : la constitution d'un groupe de travail et un co-encadrement de thèse. Notre groupe de travail a commencé par réaliser une revue de la littérature des outils permettant de visualiser des données cliniques et/ou omiques [90] afin d'identifier des pistes de recherche à explorer. La thèse d'Aarón Ayllón Benítez a débuté en 2016 sous la direction de ma collègue maître de conférences en bioinformatique, Patricia Thébault, et moi-même (co-direction que j'ai pu assurer grâce à l'obtention d'une autorisation à diriger une thèse (ADT) en 2016). Dans ce cadre, nous avons tout d'abord étudié neuf mesures de similarité sémantique afin de déterminer si l'utilisation de certaines d'entre elles étaient plus adaptées pour l'annotation de groupes de gènes [91]. Tenant compte des conclusions tirées de ce premier article, Aarón a développé un outil proposant une annotation synthétique et pertinente d'un groupe de gènes à partir de mesures de similarité sémantique et d'algorithmes qu'il a mis en œuvre et qui constitue une alternative aux méthodes d'enrichissement [92]. Par ailleurs, il a proposé une méthode de visualisation exploitant la hiérarchie d'une RTO pour parcourir de manière interactive l'annotation de plusieurs groupes de gènes [93].

Parallèlement, je dirige un autre doctorant, Jean Noël Nikiema, avec Vianney Jouhet qui est médecin de santé publique. Ma collaboration avec Vianney a débuté lors de sa thèse qui visait à rendre les données de prise en charge hospitalières exploitables pour une utilisation

secondaire de ces données en cancérologie [94]. Dans ce travail, Vianney a notamment cherché à intégrer deux RTOs diagnostiques décrivant des entités différentes : d'un côté, les diagnostics de cancer et de l'autre, la morphologie et la topographie des tumeurs [95]. Pour cela, il a conçu un modèle ontologique et décrit les RTOs dans des langages du Web sémantique. Ces représentations formelles lui ont permis d'utiliser des fonctionnalités de raisonnement pour établir des liens entre les entités distinctes décrites dans les deux RTOs étudiées. C'est dans la continuité de ce travail que nous avons co-encadré un doctorant. En effet, il nous a semblé pertinent de définir un cadre général permettant d'utiliser conjointement des RTOs décrivant des notions distinctes mais complémentaires. Jean Noël a réalisé un état de l'art des processus visant à établir des correspondances entre des entités de RTOs différentes et des méthodes mises en œuvre pour relier des entités non équivalentes [96]. Il en a conclu que l'utilisation d'une RTO de support décrite dans un langage formel était la solution optimale lorsque les RTOs à utiliser conjointement décrivent des notions distinctes. Pour étayer ses propos, Jean Noël a pu s'appuyer sur la méthode qu'il a implémentée pour intégrer les deux RTOs diagnostiques considérées précédemment dans la thèse de Vianney [57]. Il a par ailleurs proposé une méthode d'alignement entre une RTO utilisée localement par le CHU de Bordeaux et une RTO de référence décrivant des analyses biologiques. Cela a nécessité de pré-traiter les libellés de la RTO locale [97] avant de les segmenter pour comparer les mots obtenus avec ceux constituant les libellés français associés aux concepts de la RTO de référence.

Au niveau national, j'ai travaillé avec Natalia Grabar, chargée de recherche CNRS spécialiste en traitement automatique des langues. Nous avons proposé une méthode lexicale pour détecter des correspondances entre des RTOs décrivant des effets indésirables de médicaments [98]. Nous avons ensuite combiné cette méthode lexicale avec une approche basée sur le multilinguisme [99] dans le cadre du projet ANR TecSan RAVEL (Retrieval And Visualization in ELectronic health records / 2012-2014) [100]. Nous avons aussi analysé les dommages collatéraux dus à la fusion des 150 RTOs biomédicales dans le vaste système mentionné précédemment [101]. Cela nous a mené à approfondir ce travail pour comprendre pourquoi certains concepts de ce système étaient associés via plusieurs relations [102]. Nous avons déterminé si cette situation critique existait dans la RTO dont ces concepts étaient issus ou si elle était générée lors du processus de fusion. Pour prolonger cette collaboration, nous avons demandé un financement à la MESHS (Maison Européenne des Sciences de l'Homme et de la Société) de Lille pour le projet POMELO (Pathologie Médicament Alimentation / 2014-2015). Bien que représentant un financement mineur de 6000€, il nous a permis d'initier des travaux de recherche autour des connaissances liées aux médicaments avec Thierry Hamon, maître de conférences en informatique au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) et également chercheur en traitement automatique des langues. Nous nous sommes intéressés à l'interrogation des données ouvertes liées (*Linked Open Data*) grâce à des requêtes exprimées en language naturel [103, 104, 105] lors de notre participation au défi international "Question Answering over Linked Data (QALD-4)" [106]. Par ailleurs, nous avons développé un premier corpus de textes en lien avec les interactions médicament-aliment qui avaient été peu étudiées jusque-là [107]. Notre intérêt pour cette thématique émergente a résulté en l'obtention du projet ANR MIAM (Maladies, Interactions Aliments-Médicaments / 2017-2019) dont le responsable scientifique est Thierry Hamon et pour lequel j'assure la coordination des trois équipes bordelaises. Du point de vue scientifique, j'encadre le post-doctorat de Georgeta Bordea. Elle a tout d'abord mis en évidence la difficulté de retrouver des articles scientifiques sur les interactions médicament-aliment

en interrogeant PubMed de manière "classique" et testé plusieurs approches de sélection d'attributs (*features selection*) visant à déterminer les termes MeSH les plus pertinents pour trouver ces articles [108]. Elle travaille actuellement sur la modélisation des interactions médicament-aliment au travers d'une RTO formelle. Pour cela, elle réutilise une RTO qui décrit les interactions entre médicaments car celles-ci comportent des similitudes avec les interactions médicament-aliment. Il est nécessaire de l'enrichir pour prendre en compte les spécificités de notre domaine d'intérêt et d'intégrer d'autres RTOs pour représenter les connaissances liées aux aliments.

Contrairement à ce résumé de mes contributions que j'ai choisi d'organiser en fonction des collaborations que j'ai mises en place depuis ma thèse et de mes encadrements, la description détaillée de la plupart de ces travaux est articulée autour des trois problématiques introduites dans la section précédente : l'évaluation de la qualité des RTOs (chapitre 2), leur interopérabilité sémantique (chapitre 3) et la conception de nouvelles RTOs (chapitre 4). Je décris tout d'abord mes travaux sur l'évaluation de la qualité des RTOs qui est la problématique que j'ai approfondie suite à ma thèse, avec des publications principalement en première auteure. Les travaux sur l'interopérabilité et la conception de RTOs décrits dans les deux chapitres suivants sont essentiellement le fruit des encadrements que j'ai réalisés, valorisés par des articles où j'ai eu le rôle de dernière auteure ou de deuxième auteure (places alternées pour les co-encadrements). Je présente ensuite des travaux en cours sur la visualisation et l'intégration de données omiques et cliniques avant d'introduire les perspectives que j'envisage d'explorer dans les années à venir (chapitre 5). Chacun de mes travaux a été appliqué au domaine biomédical, ce qui a nécessité une appropriation conséquente en matière de connaissances métiers. C'est la raison pour laquelle je positionne mes recherches dans le champ de l'informatique médicale.

CHAPITRE 2

ÉVALUATION DE RESSOURCES TERMINO-ONTOLOGIQUES

Les travaux présentés dans ce chapitre sur l'évaluation de la qualité des RTOs visent à garantir la cohérence de leur contenu. Cette précision est nécessaire car il existe une distinction entre l'évaluation d'une RTO au regard de l'application ou, plus généralement, de l'objectif pour laquelle elle a été conçue et l'évaluation de la cohérence des éléments décrits dans la RTO. Ces deux processus distincts ont été définis dans la méthodologie METHONTOLOGY, respectivement comme la *validation* (une revue de la littérature des approches développées pour cela est décrite dans [109]) et la *vérification* [59]. C'est ce deuxième processus, également qualifié de *débogage* par Schlobach et Cornet [110] ainsi que Flouris *et al.* [111], auquel je me suis intéressée. Il se décompose en deux étapes : l'identification des incohérences et la correction de celles-ci. Dans ce cadre, les premiers travaux les plus aboutis ont été réalisés par Gómez-Pérez qui a listé des erreurs typiques au sein des hiérarchies pouvant altérer la qualité d'une ontologie [112], et Welty et Guarino qui ont défini des principes visant à identifier et supprimer des relations de subsomption si les concepts qu'elles relient sont incompatibles [113].

J'ai commencé à étudier les problématiques liées à la qualité des RTOs biomédicales pendant ma thèse et j'y ai contribué depuis en proposant des approches visant à évaluer cette qualité, et parfois à l'améliorer. Mes travaux se sont focalisés sur différentes facettes des RTOs, qui peuvent être déclinées selon les cinq facteurs suivants introduits par Zhu *et al.* dans leur revue de la littérature des approches existantes permettant d'évaluer la qualité des RTOs biomédicales [114] :

— l'**orientation conceptuelle** : garantit que les concepts ont un et un seul sens,

— la **cohérence** : analyse si des cas d'incohérence lexicale (*i.e.*, des libellés de concept contiennent le même modificateur mais les concepts ne sont pas reliés de la même façon aux concepts d'origine dont le libellé est identique, *modulo* ce modificateur) ou de classification inconsistante (*i.e.*, la classification est incohérente entre des concepts reliés hiérarchiquement) existent dans la RTO,

— la **non redondance** : s'assure qu'il n'y a pas de concepts différents représentant la même information et qu'il n'existe pas de redondance au sein des classifications (en particulier, en présence de relations transitives),

— l'**exactitude** : examine les libellés et définitions des concepts ainsi que leur classification afin d'attester que les connaissances représentées dans la RTO sont correctes,

— l'**exhaustivité de la couverture** : évalue si la RTO contient ce pour quoi elle a été définie. Ici, on cherche à détecter des erreurs telles que l'absence de certains termes, la présence de définitions seulement pour certains concepts et l'inexistence de relations entre concepts. Ce facteur contribue en partie au processus de validation susmentionné.

Ces cinq facteurs sont largement inspirés des critères définis par Gómez-Pérez pour évaluer la qualité d'une ontologie [115], complétés plus tard par Vrandečić [116], qui sont les suivants :

— la **cohérence** : garantit qu'aucune situation ne mène à des contradictions sachant que les définitions des entités sont valides. Cela regroupe les facteurs de cohérence et d'exactitude de Zhu *et al.*,

— la **complétude** : atteste qu'une ontologie contient l'ensemble des connaissances qu'elle est censée représenter, ce qui équivaut à l'exhaustivité de la couverture définie par Zhu *et al.*,

— la **concision** : veille à l'absence de redondance, correspondant à la non redondance selon Zhu *et al.*,

— l'**évolutivité** : s'assure que l'effort pour enrichir l'ontologie soit minimal et n'affecte pas le reste des connaissances,

— la **sensibilité** : évalue à quel point de légers changements dans la définition d'une entité influent sur les autres définitions établies dans l'ontologie.

Les deux derniers critères sont étroitement liés à l'évolution d'ontologies et n'apparaissent pas chez Zhu *et al.* C'est probablement dû à la large adoption du langage OWL (Web Ontology Language) [1] comme standard de description des ontologies, qui garantit leur adaptabilité [116]. En revanche, seuls Zhu *et al.* ont introduit le facteur d'orientation conceptuelle. Cela peut s'expliquer par le fait que cette problématique se pose spécifiquement pour les RTOs biomédicales dont l'unité de connaissances est parfois le terme, et non le concept.

Mes contributions quant à l'évaluation de la qualité de RTOs ont porté sur les différents critères définis par Zhu *et al.* (Tableau 2.1). J'introduis dans la section 2.1 les travaux effectués sur un vaste système intégrant plus de 150 RTOs biomédicales, l'UMLS (a,b,c), puis je présente dans la section 2.2 mes travaux pour évaluer la qualité de RTOs décrites dans un langage formel (d,e).

(a) Fleur Mougin, Olivier Bodenreider, Anita Burgun. Analyzing polysemous concepts from a clinical perspective : application to auditing concept categorization in the UMLS. Journal of Biomedical Informatics, 42(3):440 – 451, 2009

(b) Natalia Grabar, Marie Dupuch, Fleur Mougin. Dommages collatéraux de la fusion de terminologies. Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, pages 10-16, 2011

(c) Fleur Mougin, Natalia Grabar. Auditing the multiply-related concepts within the UMLS. Journal of the American Medical Informatics Association, 21(e2):e185 – 193, 2014

(d) Fleur Mougin, Olivier Bodenreider. Auditing the NCI thesaurus with semantic web technologies. AMIA Annual Symposium Proceedings, pages 500-504, 2008

(e) Fleur Mougin. Identifying redundant and missing relations in the Gene Ontology. Studies in Health Technology and Informatics, 21:195 – 199, 2015

1. `https://www.w3.org/OWL/`

	UMLS	RTO décrite dans un langage formel
Orientation conceptuelle	a,b,c	
Cohérence	a,b	d,e
Non redondance	c	e
Exactitude	b,c	
Exhaustivité de la couverture		d,e

TABLEAU 2.1 – Contributions à l'évaluation de la qualité de RTOs classées suivant les facteurs définis par Zhu *et al.* [114].

Récemment, Amith *et al.* ont complété la revue effectuée par Zhu *et al.* en classant les travaux de 2009 à 2017 visant à évaluer la qualité de RTOs biomédicales suivant les cinq facteurs précités, mais aussi suivant le type de méthodes développées [117]. Les auteurs ont recensé des méthodes structurelles, lexicales, sémantiques, basées sur un modèle d'abstraction de la RTO, basées sur des technologies liées aux données massives, sur la production participative (*crowdsourcing*), sur de la validation croisée (exploitant une ou plusieurs autres RTOs) ou encore sur des corpus. La définition des différents types de méthodes n'étant pas fournie, j'ai choisi de ne pas utiliser cette catégorisation pour présenter mes travaux. Notons cependant que trois des travaux ci-dessus apparaissent dans cette revue de la littérature, l'un ayant été classé dans les méthodes sémantiques (a) et les deux autres parmi les méthodes structurelles (c,e).

2.1 Évaluation de l'UMLS

Durant ma thèse, j'ai initié une collaboration avec Olivier Bodenreider par le biais de mon encadrante, Anita Burgun. Avec eux, nous avons travaillé sur l'évaluation de la qualité de l'UMLS (Unified Medical Language System) qui regroupe plus d'une centaine de RTOs du domaine biomédical (partie 2.1.1). C'est une ressource très riche mais qui présente de nombreuses situations problématiques [118]. Nous avons tout d'abord étudié la présence de cycles dans l'UMLS et comparé deux approches visant à les éliminer [78]. Nous avons continué nos investigations sur l'UMLS par une analyse des concepts polysémiques s'y trouvant (partie 2.1.2). Par la suite, je me suis intéressée aux concepts associés via plusieurs relations au sein de l'UMLS en collaboration avec Natalia Grabar (partie 2.1.3).

Bien que ces deux travaux se soient focalisés sur l'évaluation de caractéristiques propres à l'UMLS et non à une RTO donnée, nous précisons comment ils ont permis de mettre en évidence des incohérences au sein même de RTOs intégrées dans l'UMLS.

2.1.1 L'UMLS

L'UMLS® est une ressource créée et maintenue par la NLM qui intègre plus de 150 RTOs différentes, appelées vocabulaires sources [119] et nommées "RTOs sources" dans la suite de ce manuscrit. Cette ressource est constituée de trois composants principaux : 1) le metathesaurus®, 2) le réseau sémantique, 3) le lexique spécialisé qui contient les informations lexicales des éléments du metathesaurus nécessaires au fonctionnement d'outils de traitement automatique des langues développés par la NLM. Les méthodes proposées pour évaluer la qualité de l'UMLS étant basées sur les deux premiers composants, leur description est détaillée ci-dessous.

Le metathesaurus. C'est un large graphe qui contient plus de 3,6 millions de **concepts**, chacun d'entre eux correspondant au regroupement des termes synonymes issus des RTOs sources et étant identifié par un code unique, le *Concept Unique Identifier* ou CUI. Ces concepts sont organisés entre eux via plus de 80 millions de relations, principalement issues des RTOs sources et, pour peu d'entre elles, ajoutées par l'équipe de la NLM. Ces relations sont de différents types : *parent/child* (*PAR/CHD*) et *broader/narrower than* (*RB/RN*) correspondant aux relations hiérarchiques, les autres étant associatives.

Notons que le processus d'intégration suivi par la NLM pour constituer le metathesaurus de l'UMLS relève de la fusion (*merging*) puisqu'il contient l'ensemble des connaissances telles qu'elles sont décrites dans les RTOs sources [120].

Le réseau sémantique. C'est un réseau plus restreint composé de 133 catégories, appelées **types sémantiques** [121]. Ces types sémantiques sont reliés hiérarchiquement via des relations *is_a* mais aussi via des relations associatives elles-mêmes hiérarchisées, telles que *affects* et *complicates*. Chaque concept du metathesaurus est catégorisé par au moins un type sémantique. Les types sémantiques ont par la suite été agrégés en 15 **groupes sémantiques** qui sont disjoints deux à deux [122]. Chaque type sémantique appartient à un et un seul groupe sémantique. La figure 2.1 illustre ces différents éléments.

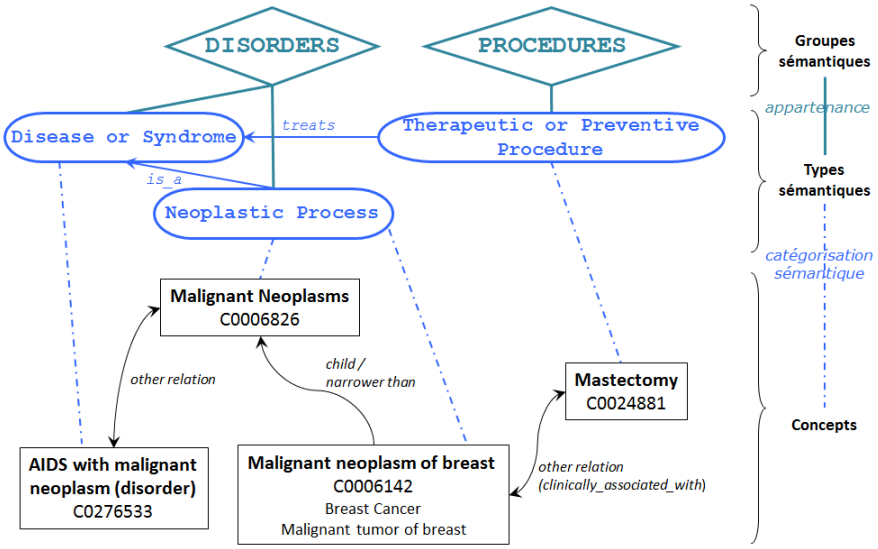

FIGURE 2.1 – Articulation entre les concepts du metathesaurus, les types sémantiques qui les catégorisent et les groupes sémantiques auxquels appartiennent les types sémantiques.

2.1.2 Polysémie dans l'UMLS

Problématique de l'ambiguïté. Dans ce travail, nous avons commencé par définir la notion de polysémie en regard de l'ambiguïté. Cette caractéristique est fréquemment rencontrée dans les différentes représentations lexicales (telles que les termes) et parfois retrouvée dans les RTOs. Les linguistes font une distinction entre l'ambiguïté contrastive et l'ambiguïté complémentaire (ou relationnelle) [123], la première correspondant à l'**homonymie** où un item lexical a deux sens bien indépendants (par exemple, la *souris* comme périphérique d'ordinateur *versus* l'animal) tandis que la deuxième concerne la **polysémie** où un item lexical a plusieurs sens mais qui sont logiquement reliés (par exemple, le *centre médical* désigne à la fois l'institution où l'on vient se faire soigner et le bâtiment qui héberge cette institution).

Dans l'UMLS, la logique voudrait que des termes ambigus appartiennent à des concepts distincts. Cependant, ce n'est pas systématiquement le cas puisque certains termes polysémiques appartiennent au même concept. C'est le cas par exemple du concept Medical center (CUI : C0565990) qui inclut les deux sens du centre médical introduits juste avant. L'UMLS traduit néanmoins le caractère polysémique de ce concept en le catégorisant avec les types sémantiques `Health Care Related Organization` et `Manufactured Object` (Figure 2.2). L'objectif de notre travail était d'étudier les concepts polysémiques de l'UMLS d'après leur catégorisation [82].

Méthodes. Comme dit précédemment, les concepts UMLS peuvent être catégorisés par plusieurs types sémantiques mais cela ne dénote pas forcément un cas de polysémie. En effet, il existe des termes à facettes multiples que l'équipe de la NLM a choisi de refléter par l'attribution de plusieurs types sémantiques. Une illustration est le concept Progesterone (C0033308) qui est catégorisé par `Steroid` (dénotant sa structure), `Hormone` et `Pharmacologic Substance` (dénotant ses différentes fonctions). Les groupes sémantiques ont par contre été créés pour représenter des sous-domaines disjoints du domaine biomédical, constituant une partition des concepts UMLS [124].

Pour étudier la polysémie dans l'UMLS, nous avons donc sélectionné les concepts qui étaient catégorisés par des types sémantiques appartenant à des groupes sémantiques différents.

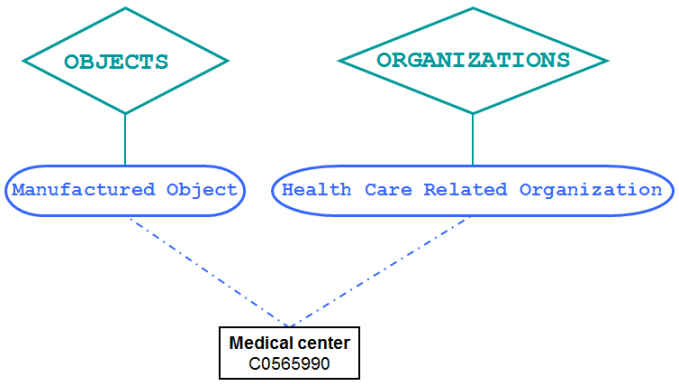

FIGURE 2.2 – Exemple de concept catégorisé par deux types sémantiques appartenant à des groupes sémantiques différents.

Une fois les concepts appartenant à plusieurs groupes sémantiques identifiés (nommés par la suite "MSG concepts" pour *multi-semantic group concepts*), nous avons analysé s'ils étaient reliés hiérarchiquement et, le cas échéant, si leur catégorisation était cohérente. Pour cela, nous nous sommes basés sur le principe d'héritage introduit par Zweigenbaum *et al.* [125], et repris par Cimino *et al.* pour étudier la catégorisation sémantique des concepts de l'UMLS [126].

> **Principe** : la catégorisation d'un MSG concept en termes de groupes sémantiques est héritée par son(ses) concept(s) parent(s).

Ce principe a deux corollaires qui précisent son application aux concepts plus génériques et spécifiques.

> **Corollaire 1** : la catégorisation d'un MSG concept en termes de groupes sémantiques est transmise à son(ses) descendant(s).

Tous les descendants d'un MSG concept doivent donc appartenir aux mêmes groupes sémantiques que ses concepts parents.

> **Corollaire 2** : la catégorisation d'un MSG concept en termes de groupes sémantiques est héritée soit par son unique concept parent, soit par ses multiples parents.

En d'autres termes, si un MSG concept n'a qu'un concept parent, les deux concepts doivent appartenir au(x) même(s) groupe(s) sémantique(s) (Figure 2.3a). Si le MSG concept a plusieurs concepts parents, il doit appartenir à chacun des groupes sémantiques auxquels appartiennent ses concepts parents (Figure 2.3b).

Pour caractériser les MSG concepts sémantiquement, nous avons constitué des clusters de MSG concepts reliés hiérarchiquement et associés à la même combinaison de groupes sémantiques. Ensuite, nous avons examiné si le principe d'héritage de la catégorisation en termes de groupes sémantiques était respecté. Pour chaque cluster, si le concept racine appartenait à la paire de groupes sémantiques SG_1-SG_2, nous avons calculé la proportion de descendants associés : (i) à la même paire de groupes sémantiques, qualifiée de **compatibilité parfaite**, (ii) à SG_1 ou à SG_2, qualifiée de **compatibilité partielle**, (iii) à au moins un groupe sémantique différent de SG_1 et SG_2, qualifiée d'**incompatibilité**.

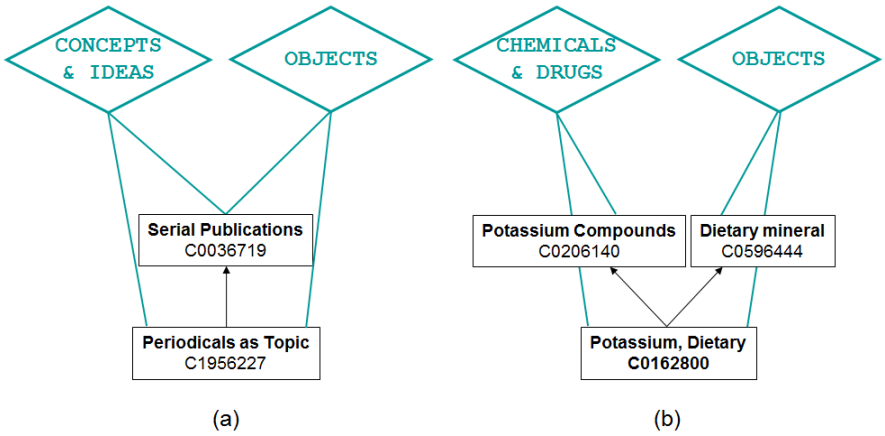

FIGURE 2.3 – Principe d'héritage au niveau des groupes sémantiques (les types sémantiques ne sont pas représentés par souci de simplification) : (a) un MSG concept appartient aux mêmes groupes sémantiques que son concept parent, (b) un MSG concept hérite d'une double catégorisation car ses concepts parents appartiennent à deux groupes sémantiques différents.

Pour compléter cette analyse quantitative, Olivier et Anita (qui sont tous deux médecins) ont examiné manuellement les MSG concepts. Ils ont considéré la catégorisation des MSG concepts et celle de leur(s) concept(s) parent(s) dans les RTOs sources et ont déterminé si l'héritage était cohérent, incohérent ou incomplet. Pour les MSG concepts dont la catégorisation était correcte, Olivier et Anita ont déterminé l'origine de la polysémie :

— par **convention** : cas où le MSG concept contient un terme déjà polysémique dans la RTO source. Par exemple, le concept Books, Illustrated (C0006003) est catégorisé dans l'UMLS par les types sémantiques `Manufactured Object` et `Intellectual Product` (qui appartiennent à deux groupes sémantiques différents) car il est défini dans le MeSH comme un matériel de support audiovisuel et un livre,

— par **intégration** : cas où les gestionnaires de l'UMLS ont choisi de regrouper dans un même concept des termes dont la définition diffère peu entre plusieurs RTOs sources car ils n'ont pas jugé la distinction suffisante pour justifier la création de plusieurs concepts. Une illustration est le concept Sesame Oil (C0036845) qui est catégorisé dans l'UMLS par les types sémantiques `Organic Chemical`, `Pharmacologic Substance` (tous deux du groupe sémantique `CHEMICALS & DRUGS`) et `Food` (du groupe sémantique `OBJECTS`) pour refléter respectivement la définition donnée par le MeSH (*The refined fixed oil obtained from the seed of one or more cultivated varieties of Sesamum indicum. It is used as a solvent and oleaginous vehicle for drugs and has been used internally as a laxative and externally as a skin softener*) et celle donnée par le NCI thesaurus [127] (*The edible oil extracted from the seeds of Sesamum indicum. Sesame oil is used as a cooking oil and as a food ingredient.*).

Résultats. Dans la version 2008AA de l'UMLS investiguée dans ce travail, nous avions identifié 1208 MSG concepts (0,08% des concepts du metathesaurus. Pour les 132 MSG concepts racines d'un cluster, le respect du principe d'héritage était le suivant :

— compatibilité parfaite pour 16 MSG concepts racines (12,1%),

— compatibilité partielle pour 63 MSG concepts racines (47,7%),

— incompatibilité pour 53 MSG concepts racines (40,2%).

Globalement, pour les 9025 descendants des 132 MSG concepts racines, 8,1% héritaient exactement des groupes sémantiques du concept racine, 64,3% en héritaient partiellement et 27,6% n'héritaient d'aucun d'entre eux.

L'analyse qualitative a montré que 91 MSG concepts (7,5%) étaient mal catégorisés. Pour les 1117 autres, la polysémie était déjà présente dans la RTO source (*i.e.*, par convention) dans 94,5% des cas.

Enfin, nous avions souligné dans ce travail que l'UMLS normalise sémantiquement les RTOs sources dans une perspective d'utilité clinique, c'est-à-dire pour être utilisé dans les systèmes d'information hospitaliers par les professionnels de santé. Pour cette raison, l'approche suivie par l'UMLS pour intégrer les RTOs sources est basée sur la création de concepts. Cela résulte généralement en une **sur-normalisation** qui génère de la polysémie mais aussi parfois en une **dé-normalisation** pour éviter que des concepts polysémiques soient intégrés à l'UMLS. Un exemple typique de la sur-normalisation est la manière dont les substances et produits pharmaceutiques décrits dans la RTO SNOMED CT [128] sont intégrés dans l'UMLS. SNOMED CT distingue ces deux notions dans deux concepts différents puisque la première dénote l'ingrédient d'un médicament tandis que l'autre correspond au médicament lui-même. Bien que cela ait un sens du point de vue ontologique, cette distinction n'a pas été reproduite dans l'UMLS [129] dont les concepteur ont choisi de catégoriser ce type de concepts par des types sémantiques de la hiérarchie `Substance` et non ceux de la hiérarchie `Clinical Drug` alors qu'il aurait été plus approprié d'effectuer une double catégorisation. En pratique, ces concepts n'apparaissent pas parmi les MSG concepts mais la hiérarchie où ils se situent n'est pas pour autant consistante (puisqu'on y trouve des concepts catégorisés comme substances et d'autres comme produits). À l'inverse, la dé-normalisation correspond à la volonté de l'UMLS de décomposer des concepts qui sont agrégés dans les RTOs sources. C'est par exemple le cas du MeSH qui regroupe des concepts différents dans une même catégorie, ce qui est utile pour des besoins de recherche d'information mais incompatible avec le principe d'utilité clinique motivant les concepteurs de l'UMLS.

Selon la classification de Zhu *et al.*, l'aspect lié à la polysémie étudié dans ce travail relève du facteur d'orientation conceptuelle tandis que l'étude de la catégorisation des concepts polysémiques au regard du principe d'héritage aborde la question de cohérence au sein de l'UMLS.

2.1.3 Concepts reliés de manière multiple dans l'UMLS

Problématique. Lors de l'intégration des différentes RTOs sources dans l'UMLS, les termes issus de chacune d'entre elles sont inclus dans des concepts et l'ensemble des relations existant entre ces termes sont également ajoutées. Dans un travail préliminaire réalisé avec Natalia Grabar, nous avons recensé les problèmes posés par la fusion des RTOs sources au sein de l'UMLS [101]. En particulier, nous avons identifié que cela peut résulter en des relations multiples, parfois contradictoires, entre deux concepts. Le travail présenté dans cette partie est une étude de ces relations multiples visant à comprendre la raison de leur existence [102].

Méthodes. Dans la version 2012AA de l'UMLS analysée dans ce travail, il existait onze relations actives que nous avons classées en trois grandes catégories : les relations de synonymie, les relations hiérarchiques et les relations associatives (Tableau 2.2). Il y avait environ 300 relations issues des RTOs sources qui ont été assignées à l'une de ces onze relations actives par les concepteurs de l'UMLS. Par exemple, la relation d'origine [2] *same_as* a été assignée à la relation active *SY*, *inverse_isa* à *PAR* et *has_component* à *RO*.

Catégorie	Abréviation	Signification
Synonymie	SY	source asserted synonymy
Hiérarchique	CHD	has child relationship
	PAR	has parent relationship
	RB	has a broader relationship
	RN	has a narrower relationship
	SIB	has sibling relationship
Associative	AQ	allowed qualifier
	QB	can be qualified by
	RO	has relationship other than synonymous, narrower or broader
	RL	has similar or "alike" relationship
	RQ	related and possibly synonymous

Tableau 2.2 – Relations UMLS et catégorie à laquelle elles appartiennent

Avant d'analyser les relations multiples, nous avons tout d'abord gardé une seule occurrence de chaque relation par paire de concepts. Une fois ce filtrage fait, nous avons examiné la compatibilité des relations multiples existant entre chaque paire de concepts (C_1,C_2) en distinguant quatre cas (Figure 2.4) :

(a) combinaisons **contradictoires** : cas où des relations inverses sont rencontrées parmi les multiples relations (*e.g.*, *CHD PAR*, *RB RL RN*, *AQ QB*, *PAR RN RO SIB*). Notons que nous avons considéré que *PAR* et *RN* ainsi que *CHD* et *RB* étaient contradictoires,

(b) combinaisons impliquant des relations de **granularité différente** : situations où SIB et/ou SY sont combinées à au moins une des relations hiérarchiques *PAR*, *CHD*, *RB* et *RN* (*e.g.*, *PAR SIB*, *PAR RO SIB*, *RB SY*),

(c) combinaisons **hétérogènes** : combinaisons impliquant des relations de catégories différentes (*i.e.*, synonymie, hiérarchique et associative) (*e.g.*, *PAR RO*, *RQ SIB*, *PAR RB RO RQ*),

(d) combinaisons **homogènes** : combinaisons *PAR RB* et *CHD RN* ou toute combinaison de relations associatives - hormis *AQ QB* car ce sont deux relations inverses (*e.g.*, *PAR RB*, *RL QB*, *AQ RO RQ*).

Nous avons associé les paires de concepts à un seul type de combinaisons, dans l'ordre de présentation ci-dessus.

2. La notion de *relation d'origine* est utilisée par la suite pour dénoter une relation telle qu'elle existe dans une RTO source, par opposition aux relations actives de l'UMLS

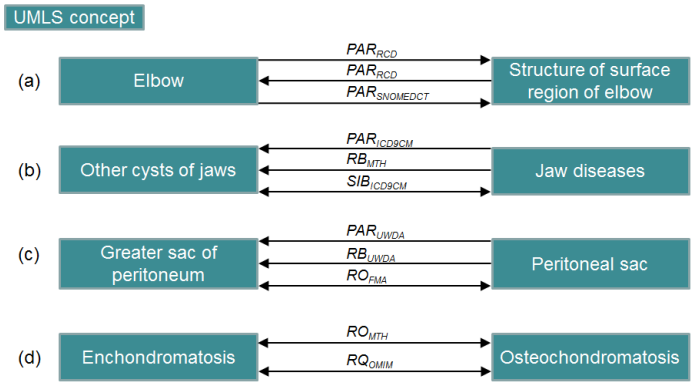

FIGURE 2.4 – Les différents types de combinaisons entre des concepts UMLS reliés par plusieurs relations : (a) contradictoires, (b) de granularité différente, (c) hétérogènes, (d) homogènes.

Nous avons ensuite cherché à déterminer la raison pour laquelle des relations multiples existaient entre deux concepts. Cette analyse a été réalisée automatiquement puisque l'UMLS précise l'origine de chaque relation. Dans les situations où les relations multiples entre deux concepts étaient issues de RTOs sources différentes, nous avons conclu que c'était dû au processus d'intégration dans l'UMLS (*i.e.*, par **intégration** si l'on reprend le vocabulaire utilisé dans la partie 2.1.2). Par contre, lorsque la combinaison de relations existait dans la RTO source, nous avons examiné si c'était une même paire de termes qui étaient reliés plusieurs fois au sein de cette RTO. Le cas échéant, la combinaison a été étiquetée par **convention** alors que dans le cas contraire, elle a été qualifiée par **intégration**.

Résultats. Notre étude s'est focalisée sur 439 087 paires de concepts reliés de manière multiple (soit 3,6% du nombre total de paires de concepts reliés dans l'UMLS). Le détail des résultats de la compatibilité et de l'origine des combinaisons est présenté dans la figure 2.5.

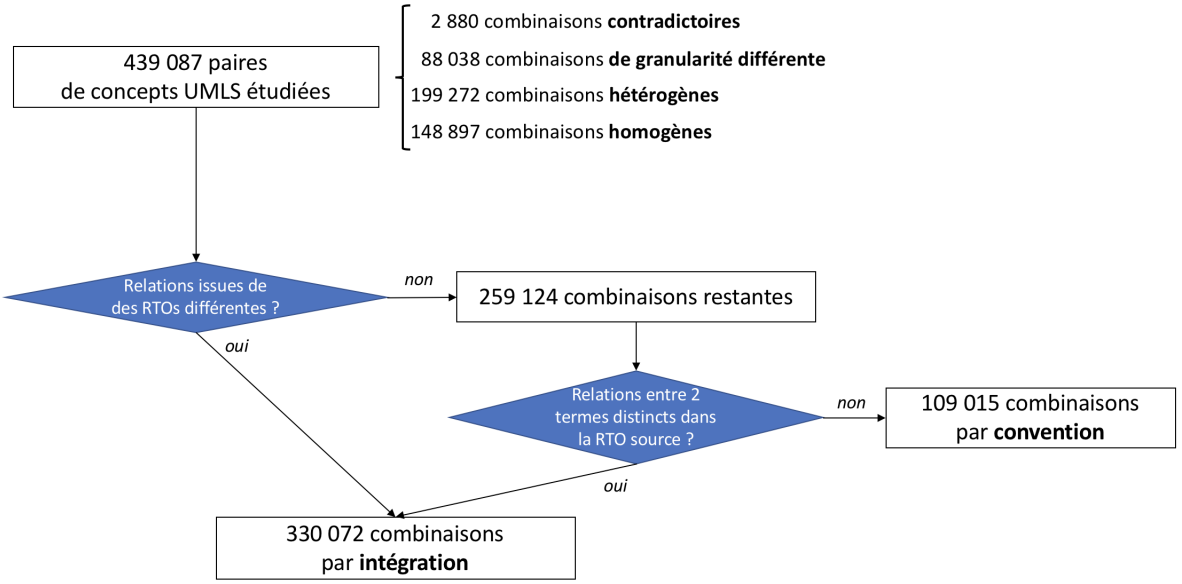

FIGURE 2.5 – Nombre de combinaisons de concepts UMLS étudiées et répartition en terme de compatibilité (accolade) et en fonction de l'origine de la combinaison (arbre de décision).

En terme de compatibilité, nous avons trouvé que 0,7% des combinaisons étaient contradictoires et 20% des combinaisons impliquaient des relations de granularité différente. Ce sont les combinaisons hétérogènes qui étaient les plus fréquentes avec 45,4%.

L'origine des combinaisons était :

— par **intégration** pour 75,2% des paires de concepts. Plus précisément, la combinaison de relations n'existait pas au préalable dans une RTO source pour 41% des paires de concepts (Figure 2.6a). Pour les 34,2% des paires de concepts restantes, une RTO source reliait de manière multiple les concepts mais les relations impliquées associaient des termes distincts dans cette RTO (Figure 2.6b),

— par **convention** pour 24,8% des paires de concepts. L'examen de ce type de situation a révélé que ces combinaisons étaient souvent homogènes et reflétaient plutôt des points de vue différents mais cohérents permettant d'exprimer le lien entre deux termes. Néanmoins, nous avons observé que certaines RTOs sources décrivaient des relations non homogènes entre deux termes (Figure 2.6c), ce qui nous a permis d'identifier la présence de relations redondantes, inappropriées, ou encore sous-spécifiées dans ces RTOs.

Notons que notre analyse au niveau terme peut avoir sous-estimé la responsabilité des RTOs sources quant à la présence de relations multiples entre concepts. MedDRA est une bonne illustration de ce type de RTOs : elle utilise des termes différents pour représenter un concept unique [130], résultant en une situation telle que celle observée en Figure 2.6b alors que c'est la RTO source qui est responsable de la combinaison de relations observée dans ce genre de cas (puisqu'il est logique que l'UMLS regroupe les deux termes au sein du même concept).

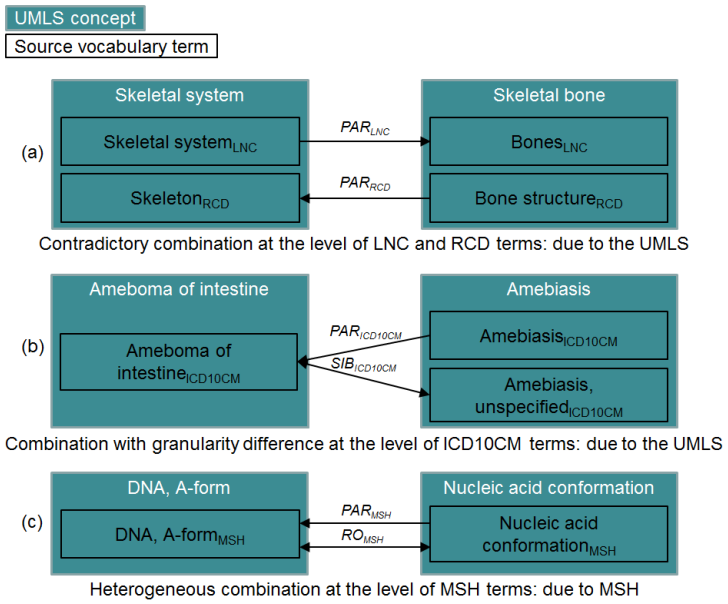

FIGURE 2.6 – Illustrations de l'existence de relations multiples entre concepts UMLS (source : [102]) : (a) une combinaison contradictoire générée au moment de l'intégration des RTOs LOINC (LNC) et Read Codes (RCD) dans l'UMLS, (b) une combinaison de relations de granularité différente également générée au moment de l'intégration de la CIM10 modifiée (ICD10CM, CM pour *clinical modification*) dans l'UMLS, (c) une combinaison hétérogène pré-existant dans le MeSH (MSH).

Finalement, nous avons analysé à la main 288 combinaisons contradictoires (10% sélectionnées aléatoirement) car ce sont celles qui étaient les plus problématiques. Nous avons observé que le

processus d'intégration dans l'UMLS était responsable de leur existence dans 99,9% des cas. Ces combinaisons contradictoires étaient dues à diverses raisons de nature linguistique, telles que :

— La valeur sémantique des termes composés (*e.g.*, colorectal et colon/rectum), étudiés en détail dans [131], et des termes reliés par des conjonctions de coordination (*e.g.*, et, ou) peut être différente d'une RTO source à l'autre. Par exemple, dans MedDRA, Esophageal stenosis (C0014866) est plus général que Oesophageal stenosis and obstruction (C0851721) tandis que la relation est inverse dans la RTO MEDCIN. Nous avons supposé que cela venait de la signification attribuée à la coordination : certaines RTOs l'utilisent pour créer des termes plus génériques, tandis que d'autres l'utilisent pour spécifier les termes (et donc en créer de plus spécifiques).

— La nature implicite de certains modificateurs peut impacter les relations entre concepts. Par exemple, Total nephrectomy (C0176996) et Nephrectomy (C0027695) étaient associés via cinq relations (*PAR, RN, RO, SIB, SY*). C'est le cas où le modificateur *total* est considéré comme implicite qui explique la présence de la relation *SY*.

— Les liens fonctionnels et causaux entre termes peuvent présenter de grandes variations lorsqu'ils sont représentés via des relations hiérarchiques. Par exemple, Angioedema (C0002994) et Urticaria (C0042109), qui sont des manifestations communes de réactions allergiques, étaient reliés par *CHD, PAR, RB, RO, RQ* et *SIB*.

Dans ce travail, nous avons analysé non seulement l'orientation conceptuelle selon la classification de Zhu *et al.* en examinant les termes regroupés dans un même concept UMLS mais aussi la non redondance grâce à l'identification de relations redondantes et l'exactitude en détectant des relations inappropriées parmi celles que nous avons étudiées.

2.1.4 Conclusions

Ces deux études ont permis de mettre en évidence des situations potentiellement problématiques générées par l'intégration de multiples RTOs au sein d'un système unique qu'est l'UMLS. Le facteur majeur de Zhu *et al.* affecté dans ce cadre est l'orientation conceptuelle. Plus précisément, nous avons analysé les concepts polysémiques et les relations multiples entre concepts avec pour objectif d'expliquer leur présence dans l'UMLS. Nous avons développé des approches automatiques pour faciliter cette interprétation puis nous avons réalisé une analyse manuelle pour mettre en lumière des situations typiques engendrant ce type de problèmes. Par ailleurs, ces travaux ont apporté des conclusions additionnelles et complémentaires à des travaux passés, en particulier celui qui visait à analyser les concepts catégorisés par plusieurs types sémantiques [132] et celui qui a étudié les relations multiples entre concepts UMLS mais en se limitant à une analyse manuelle et non systématique de celles-ci [133].

Cependant, dans les deux cas, notre analyse s'est focalisée sur un nombre limité de ces situations. Dans la première étude, les concepts polysémiques ont été détectés de par leur appartenance à plusieurs groupes sémantiques, mais d'autres cas de polysémie existant dans l'UMLS n'ont pas été examinés. Pour les relations multiples, nous avons étudié celles qui existent entre des concepts distincts alors que des concepts UMLS sont reliés de manière multiple à eux-mêmes. Notons par ailleurs que dans l'UMLS se trouvent certaines RTOs sources dont la hiérarchie n'est pas basée sur la relation *is_a* telles que le MeSH, ce qui peut poser problème quant à nos hypothèses sur le principe d'héritage ou encore sur notre caractérisation des combinaisons de relations multiples.

2.2 Évaluation de ressources termino-ontologiques décrites dans un langage formel

Au cours de ma thèse, j'ai utilisé des technologies du Web sémantique pour concevoir un système d'intégration. Dans la continuité du premier travail réalisé avec Olivier Bodenreider, nous avons étudié l'intérêt de disposer d'une RTO décrite dans un langage du Web sémantique pour évaluer sa qualité. Plus précisément, nous avons analysé la qualité des relations présentes dans le NCI thesaurus que nous avons préalablement représenté en RDF (partie 2.2.1). Dans un deuxième travail, j'ai proposé une méthode permettant d'évaluer et d'améliorer la description formelle de la Gene Ontology (partie 2.2.2).

2.2.1 Qualité des relations du NCI thesaurus

Problématique. L'automatisation des méthodes d'évaluation de la qualité des RTOs repose généralement sur le développement *ad hoc* de programmes informatiques, comme dans les deux travaux présentés à la section précédente. Dans le travail présenté ci-après, nous avons étudié la possibilité de tirer profit des technologies du Web sémantique pour effectuer ce type d'évaluation [81]. Plus précisément, notre objectif a été d'évaluer la qualité des relations hiérarchiques et associatives du NCI thesaurus décrit en RDF (Resource Description Framework) [3] au moyen de requêtes SPARQL (SPARQL Protocol and RDF Query Language) [4].

Le NCI thesaurus. Le NCI thesaurus (NCIt) est une RTO de référence développée par le National Cancer Institute américain, qui fournit une large couverture du domaine de la cancérologie [134]. Dans la version 2007_05E que nous avons utilisée lors de cette étude, le NCIt contenait quasiment 60 000 concepts. Plus de 90 types de relations y étaient définies, comme par exemple *disease_has_abnormal_cell* pour qualifier le lien entre une pathologie et un type de cellule. Le NCIt est disponible dans plusieurs formats, y compris OWL.

Méthodes. Dans ce travail, nous avons procédé en deux étapes : 1) conversion du NCIt en triplets et intégration dans une base de données de type graphe permettant de stocker uniquement des triplets, nommée *triple store* [5], et ajout de triplets contenant des informations issues de l'UMLS, en particulier les types et groupes sémantiques, 2) requêtes SPARQL sur le *triple store* pour évaluer la qualité des relations hiérarchiques et associatives entre les concepts du NCIt.

Pour chaque concept du NCIt, nous avons extrait du fichier OWL son code, son **terme préféré** (*i.e.*, le terme à utiliser prioritairement s'il en existe plusieurs pour désigner le concept) et ses concepts parents. À partir de ces informations, nous avons créé des triplets RDF dont le sujet était le concept (désigné par son code NCIt), le prédicat était la relation à laquelle le concept participait et l'objet était le concept ou le terme qui lui étaient associés. En plus de ces triplets, nous avons ajouté le CUI auquel appartenait le code NCIt dans l'UMLS ainsi que les types sémantiques le catégorisant dans l'UMLS. Par souci de simplicité, les relations associatives entre les concepts du NCIt ont également été récupérées dans l'UMLS. Enfin, l'ensemble des

3. `https://www.w3.org/RDF/`
4. `https://www.w3.org/2001/sw/wiki/SPARQL`
5. Nous avons utilisé le *triple store* Mulgara™(`http://mulgara.org`)

types et groupes sémantiques de l'UMLS ainsi que les relations entre types sémantiques et entre relations ont été intégrés au *triple store*. Une fois le *triple store* créé, nous avons pu l'interroger via des requêtes SPARQL afin d'effectuer l'évaluation.

Les relations hiérarchiques ont été examinées au niveau des types sémantiques et au niveau des groupes sémantiques, d'après des principes d'héritage similaires à ceux utilisés dans la partie 2.1.2. Pour chaque paire de concepts (C_1,C_2) reliés hiérarchiquement, où C_1 est un concept enfant de C_2, nous avons étudié leur type sémantique, respectivement ST_1 et ST_2, et vérifié si ST_1 était le même type sémantique ou un descendant de ST_2. Lorsque les concepts étaient catégorisés par plusieurs types sémantiques, tous ceux catégorisant C_1 devaient être les mêmes ou des descendants d'au moins un des types sémantiques catégorisant C_2 (Figure 2.7a). Au niveau des groupes sémantiques, nous avons simplement vérifié si les types sémantiques catégorisant C_1 et C_2 appartenaient au même groupe sémantique (Figure 2.7b).

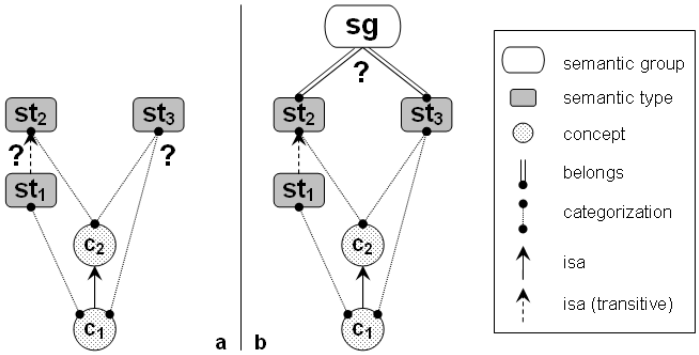

FIGURE 2.7 – Évaluation des relations hiérarchiques (source : [81]) : (a) au niveau des types sémantiques, (b) au niveau des groupes sémantiques.

La requête suivante nous a permis de récupérer les concepts dont la catégorisation sémantique au sein de l'UMLS était incompatible avec celle de leur concept parent (la clause *minus* effectue une différence ensembliste pour ignorer les concepts dont le type sémantique est le même ou un descendant (grâce à la clause *trans* qui active le caractère transitif de la relation *subClassOf*) du type sémantique de son concept parent) :

```
select $c1 from <rmi://localhost/server1#modelncit>
where (
      ($c1 <NCIt:hasUMLSST> $st1)
      and ($c2 <NCIt:hasUMLSST> $st2)
      and ($c1 <rdfs:subClassOf> $c2)
)
minus (
      ($c1 <NCIt:hasUMLSST> $st1)
      and ($c2 <NCIt:hasUMLSST> $st2)
      and ($c1 <rdfs:subClassOf> $c2)
      and (
            ($c2 <NCIt:hasUMLSST> $st1)
            or ($st1 <rdfs:subClassOf> $st2)
            or trans($st1 <rdfs:subClassOf> $st2)
      )
)
```

Pour évaluer la qualité des relations associatives, nous voulions comparer celles-si avec les relations existant entre les types sémantiques catégorisant les concepts reliés. Comme il n'existait pas de correspondance entre les relations du NCIt et celles du réseau sémantique, nous en avons proposé pour 19 relations du NCIt reliant des concepts de type "Disease" à d'autres concepts. Pour cela, nous avons essayé de trouver une correspondance entre les domaines et co-domaines (*i.e.*, *domain* et *range* utilisés dans le langage OWL) décrits pour les relations du NCIt et les types sémantiques de l'UMLS. Lorsque nous y sommes parvenus, nous avons examiné les relations existant dans le réseau sémantique entre ces types sémantiques et regardé plusieurs paires de concepts catégorisés par ces derniers pour évaluer la possible correspondance. À la fin de ce processus, chacune des 19 relations du NCIt a été mise en correspondance (via une équivalence ou une relation de subsomption) avec une relation du réseau sémantique. Ensuite, la qualité des relations associatives a été évaluée comme illustré en figure 2.8. Ainsi, une relation r_{nci} entre deux concepts C_1 et C_2 était considérée correcte si cette relation était équivalente ou plus spécifique que la relation r_{sn} existant dans le réseau sémantique entre ST_1 et ST_2 catégorisant C_1 et C_2, respectivement. Notons que si ces concepts étaient catégorisés par plus d'un type sémantique, la relation devait être cohérente avec toutes les relations existant entre les types sémantiques correspondants.

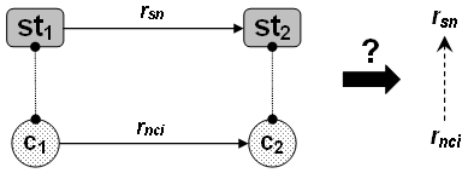

FIGURE 2.8 – Évaluation des relations associatives (source : [81]).

Un des avantages d'utiliser un *triple store* est qu'il est possible d'effectuer du raisonnement. Nous avons ainsi créé des règles d'inférence pour implémenter la réflexivité et la transitivité de la relation *subPropertyOf* définie entre des relations associatives du NCIt et entre les relations du réseau sémantique ainsi que la disjonction entre les groupes sémantiques via une relation *isDisjoint*.

Résultats. Sur les 58 843 concepts du NCIt, 30% présentaient une incompatibilité au niveau des types sémantiques et 11% au niveau des groupes sémantiques. En ce qui concerne les relations associatives, la cohérence était globalement bonne. Le détail des résultats est donné dans le tableau 2.3. Pour huit types de relations, aucune incohérence n'a été trouvée et pour trois d'entre elles, la cohérence n'a pas pu être vérifiée car la relation correspondante dans le réseau sémantique ne reliait pas les types sémantiques associés aux domaine et co-domaine des relations du NCIt. C'était le cas par exemple de la relation *disease_is_stage*. L'UMLS catégorise les grades d'une maladie (*i.e.*, les concepts cibles de cette relation) comme `Intellectual Product` et ce type sémantique n'est pas relié à `Disease or Syndrome` (*i.e.*, catégorisant les concepts sources de cette relation). Pour les huit relations restantes, des incohérences ont été identifiées et, en examinant certaines d'entre elles en détail, nous avons constaté qu'elles étaient essentiellement dues à des erreurs de catégorisation ou à l'absence de relations décrites dans le réseau sémantique.

Relation dans le NCIt	Relation correspondante dans le réseau sémantique	Nb de relations dans l'UMLS	Nb d'incohérences d'après la catégorisation UMLS
disease_has_abnormal_cell	has_location	7085	8
disease_has_associated_anatomic_site	has_location / produces	6208	301
disease_has_associated_disease	co-occurs_with / occurs_in	368	2
disease_has_cytogenetic_abnormality	has_result	124	0
disease_has_finding	has_manifestation	5960	39
disease_has_metastatic_anatomic_site	has_location	89	2
disease_has_molecular_abnormality	has_result	451	0
disease_has_normal_cell_origin	has_location	5913	0
disease_has_normal_tissue_origin	has_location / produces	6325	78
disease_has_primary_anatomic_site	has_location / produces	4831	250
disease_is_stage	has_result	∅	∅
disease_is_grade	has_evaluation	∅	∅
eo_disease_has_associated_cell_type	has_location	13	0
eo_disease_has_property_or_attribute	has_property	∅	∅
eo_disease_has_associated_eo_anatomy	has_location / produces	1313	0
eo_disease_maps_to_human_disease	co-occurs_with	735	4
gene_associated_with_disease	location_of	1017	0
gene_product_malfunction_associated_with_disease	causes / produced_by	242	0
regimen_has_accepted_use_for_disease	treats / results_of	229	0

TABLEAU 2.3 – Pour chaque type de relation associative du NCIt et la(les) relation(s) correspondante(s) (équivalente ou plus générique) dans le réseau sémantique, sont donnés : le nombre de relations entre des concepts catégorisés par l'UMLS et le nombre de relations incohérentes d'après la catégorisation UMLS.

En résumé, ce travail s'est intéressé au critère de cohérence défini par Zhu *et al.* en se basant sur les principes d'héritage de la catégorisation des concepts du NCIt. De plus, le dernier constat concernant le manque de certaines relations relève du facteur d'exhaustivité de la couverture.

2.2.2 Relations manquantes et redondantes dans la Gene Ontology

Problématique. Dans le travail présenté ci-après, j'ai axé l'évaluation sur la détection de relations redondantes mais surtout sur l'identification de relations manquantes au sein de la Gene Ontology [135]. La première situation pose problème dans le sens où les relations redondantes ne sont pas nécessaires (par définition) alors qu'elles peuvent contribuer à la génération d'un graphe enchevêtré inutilement complexe à manipuler. Dans le deuxième cas, l'absence de relations qui devraient être décrites peut être un frein aux raisonnements pouvant être faits sur la RTO considérée et compliquer sa maintenance, notamment.

La Gene Ontology. La Gene Ontology (GO) est une RTO fournissant un vocabulaire contrôlé et commun pour décrire le rôle des gènes et produits de gènes de n'importe quel organisme [36]. Elle est structurée sous la forme de trois hiérarchies distinctes contenant plus de 25 000 processus biologiques, 9 600 fonctions moléculaires et 3 400 composants cellulaires. Les concepts GO (nommés *termes* par les concepteurs de la GO) sont organisés suivant la relation *is_a* et les relations associatives *part_of*, *regulates*, *positively_regulates* et *negatively_regulates* établissent des liens entre des concepts de hiérarchies différentes.

La GO est disponible dans plusieurs formats, notamment OWL et OBO (Open Biomedical Ontologies) [6] qui sont tous les deux des langages formels. J'ai choisi d'étudier le fichier *go.obo* [7] car OBO est le format natif de la GO et parce qu'il contient des relations supplémentaires (que sont *has_part*, *occurs_in* et *results_in*). Des **axiomes**, qui sont des énoncés considérés comme vrais et sur lesquels on peut raisonner, permettent d'attribuer des définitions logiques à certains concepts (via les tags *intersection_of*). Par exemple, le concept regulation of T cell activation est défini comme suit dans le langage OBO :

```
id: GO:0050863
name: regulation of T cell activation
is_a: GO:0051249 ! regulation of lymphocyte activation
intersection_of: GO:0065007 ! biological regulation
intersection_of: regulates GO:0042110 ! T cell activation
```

La définition logique correspondante en logique de description [25] est la suivante (dans le langage OBO, les relations associatives sont décrites par défaut avec le quantificateur existentiel) :

$$\mathsf{RegulationOfTCellActivation} \equiv \mathsf{BiologicalRegulation} \sqcap \exists\, \mathit{regulates}.\mathsf{TCellActivation}$$

Méthodes. Pour détecter les relations redondantes, j'ai exploité les règles de composition définies par les gestionnaires de la GO [8] pour 16 combinaisons de relations. J'ai donc vérifié automatiquement si des relations décrites entre des concepts GO pouvaient être inférées d'après ces combinaisons. Le cas échéant, je les ai considérées comme redondantes.

Pour les relations manquantes, j'ai exploité la compositionnalité des termes de la GO [136, 137]. Ainsi, j'ai extrait tous les bi-grammes (*i.e.*, deux mots consécutifs) existant au sein des termes préférés des concepts GO. J'ai alors sélectionné les bi-grammes les plus fréquents et étudié ceux qui apparaissaient plus de 1000 fois. Ensuite, j'ai examiné manuellement les définitions

6. http://owlcollab.github.io/oboformat/doc/GO.format.obo-1_2.html
7. http://snapshot.geneontology.org/ontology/go.obo
8. http://geneontology.org/page/ontology-relations#summary

logiques des concepts GO dont les termes préférés contenaient ces bi-grammes. Lorsqu'une règle basée sur la compositionnalité de ces termes pouvait être établie, je l'ai utilisée afin de vérifier automatiquement que tous les concepts dont le terme préféré avait la même structure disposait d'une définition logique de la forme correspondante. Lorsque ça n'était pas le cas, une relation manquante pouvait ainsi être identifiée. Par exemple, pour les concepts dont le terme préféré a comme structure compositionnelle regulation of X, la définition logique suivante pouvait être proposée (considérant que `X` est le terme préféré d'un concept GO) :

$$\mathsf{RegulationOfX} \equiv \mathsf{BiologicalRegulation} \sqcap \exists\, \mathit{regulates}.\mathsf{X}$$

Résultats. Au total, 1041 relations redondantes ont été identifiées (Tableau 2.4). Parmi les 16 combinaisons pour lesquelles de l'inférence était possible, seules 10 d'entre elles présentaient des relations redondantes. Un exemple d'une telle relation est donnée en figure 2.9.

Combinaison de relation	Relation inférée	Nb de relations redondantes
is_a * is_a	is_a	804
is_a * part_of	part_of	187
part_of * part_of	part_of	23
part_of * is_a	part_of	11
is_a * positively_regulates	positively_regulates	6
is_a * negatively_regulates	negatively_regulates	4
is_a * regulates	regulates	3
has_part * is_a	has_part	1
is_a * has_part	has_part	1
has_part * has_part	has_part	1

TABLEAU 2.4 – Nombre de relations redondantes retrouvées dans la Gene Ontology pour chaque combinaison.

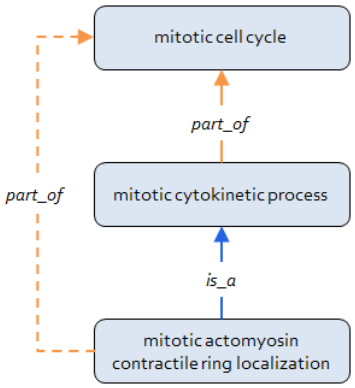

FIGURE 2.9 – Exemple d'une relation redondante (en pointillés) existant entre des concepts de la Gene Ontology.

Neuf bi-grammes apparaissaient dans plus de 1000 termes préférés des concepts GO (Tableau 2.5).

Bi-gramme	Fréquence
regulation of*	8870
positive regulation*	2898
negative regulation*	2845
biosynthetic process	2189
involved in*	2029
metabolic process	1771
catabolic process	1545
response to	1521
signaling pathway	1048

TABLEAU 2.5 – Bi-grammes les plus fréquents parmi les termes préférés des concepts de la Gene Ontology. Les bi-grammes ayant été étudiés dans ce travail sont suivis d'une astérisque.

J'ai centré mon analyse sur les bi-grammes "regulation of", "positive regulation", "negative regulation" et "involved in", les autres bi-grammes ayant déjà été étudiés par Ogren *et al.* [136]. Pour 7559 des 8870 concepts GO de la forme (positive/negative) regulation of X (85,2%), le terme préféré X existait. Pour 28 d'entre eux, j'ai identifié une définition logique manquante (Figure 2.10a). En regardant les cas où X n'existait pas, j'ai identifié des formes dérivées dont la plus commune était (positive/negative) regulation of X by Y. Parmi les 323 concepts dont le terme préféré était de cette forme, 149 d'entre eux ne disposait pas de la définition logique suivante (exemple Figure 2.10b) :

$$\text{(Positive/Negative)RegulationOfXByY} \equiv \text{Y} \sqcap \exists\, \textit{results\_in}.\text{(Positive/Negative)RegulationOfX}$$

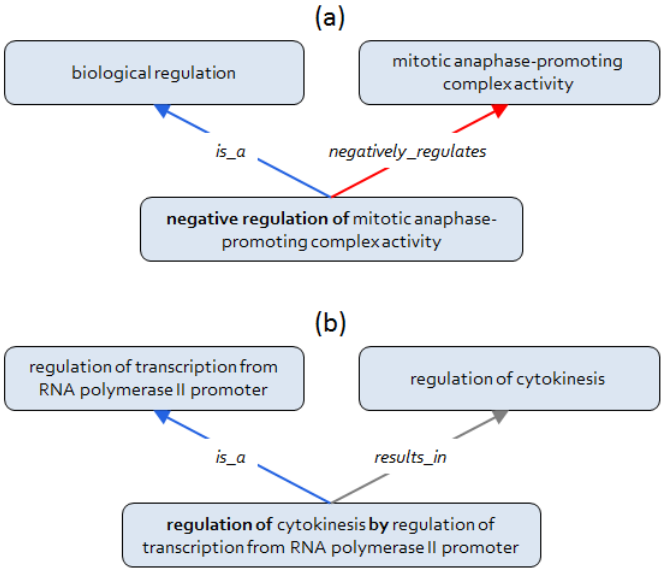

FIGURE 2.10 – Exemple de relations manquantes entre concepts dans la Gene Ontology et dont le terme préféré a pour structure compositionnelle : (a) negative regulation of X, (b) regulation of X by Y.

J'ai ensuite considéré les 2029 concepts GO dont le terme préféré était de la forme `X involved in Y`. Pour 941 d'entre eux, les termes `X` et/ou `Y` n'existaient pas. Pour les autres concepts, 765 d'entre eux avaient la définition logique $\text{X} \sqcap \exists\, \textit{part\_of}.\text{Y}$. J'ai alors identifié que cette définition était manquante pour les 323 concepts restants. Par exemple, la définition suivante a été identifiée comme manquante :

$$\text{SomaticHypermutationOfImmunoglobulinGenesInvolvedInImmuneResponse} \equiv \text{SomaticHypermutationOfImmunoglobulinGenes} \sqcap \exists\, \textit{part\_of}.\text{ImmuneResponse}$$

Dans ce travail, j'ai étudié le facteur de cohérence défini par *Zhu et al.* avec un focus sur la facette lexicale des concepts GO, contrairement aux autres travaux présentés dans ce chapitre qui ont identifié des incohérences au regard de la catégorisation sémantique des concepts décrits dans les RTOs. Les critères de non redondance et d'exhaustivité de la couverture ont aussi été abordés par la découverte de relations redondantes et manquantes.

2.2.3 Conclusions

Les deux travaux présentés dans cette section ont permis d'évaluer l'exhaustivité de la couverture offerte par la RTO étudiée, ce qui n'est pas le cas de ceux décrits dans la section précédente. Notons que dans l'étude sur la GO, la détection de relations manquantes a pu être automatisée, même si l'identification du patron syntaxique fréquemment utilisé pour un ensemble de termes ayant la même structure compositionnelle a été faite manuellement.

L'étude effectuée sur le NCIt a permis de montrer l'intérêt d'utiliser des langages formels pour réaliser l'évaluation de RTOs sans développer de programme *ad hoc*. Nous avons choisi de faire cette étude via des requêtes SPARQL posées sur un *triple store* car, en 2008, les raisonneurs OWL étaient assez limités en termes de performance sur de larges RTOs. Un tel raisonneur aurait probablement pu être utilisé pour évaluer la qualité du NCIt mais cela n'aurait certainement pas fonctionné pour de plus vastes RTOs telles que la SNOMED CT. Pour cette raison, nous avons proposé une approche basée sur RDF et SPARQL, qui a effectivement été utilisée avec succès sur la SNOMED CT en 2010 [138]. Notons cependant que, depuis, les performances des raisonneurs OWL se sont largement améliorées et d'autres raisonneurs tels que ELK [139] ont été développés afin de traiter des RTOs de grande taille en limitant les raisonnements au quantificateur existentiel (qui sont suffisants pour l'étude que nous avons réalisée sur le NCIt).

L'ajout d'axiomes à la GO à partir des termes GO et des définitions textuelles a été relativement tardif mais a fait l'objet de multiples efforts pour associer des définitions logiques aux concepts GO [140, 141, 142, 143, 144]. Le travail présenté ici a permis de mettre en évidence que certaines relations manquaient encore pour que chaque concept GO ait une définition logique. Des travaux plus récents ont proposé des approches plus systématiques pour identifier des treillis de concepts (où chaque paire de concepts n'a qu'un seul ancêtre commun minimal, à savoir le plus spécifique) grâce aux similarités lexicales de leurs termes préférés [145]. Les auteurs ont détecté dans un deuxième temps des relations manquantes et des concepts manquants, comme ils l'avaient déjà fait dans la SNOMED CT [146] et le NCIt [147]. Notons que les treillis de concepts avaient été utilisés il y a environ 10 ans dans le même objectif, mais en exploitant les définitions logiques des concepts [148, 149].

2.3 Synthèse et perspectives

Dans ce chapitre, nous avons fait trois **constats** : (i) l'UMLS est une ressource qui présente de nombreuses incohérences, (ii) la disponibilité d'une RTO dans un langage formel influe sur l'évaluation de sa qualité, et (iii) même si l'évaluation de la qualité d'une RTO peut être automatisée, une intervention manuelle est généralement nécessaire pour comprendre la présence des erreurs et effectuer les corrections appropriées au sein de la RTO.

L'UMLS est une ressource très riche car elle intègre la plupart des RTOs les plus utilisées dans le domaine biomédical et fournit un modèle de connaissances de haut niveau grâce aux types sémantiques [121]. Cependant, la première section de ce chapitre a montré que son approche d'intégration pose de nombreux problèmes en terme de qualité. D'une part, le fait de regrouper les termes issus des RTOs sources au sein d'un même concept génère des erreurs liées aux facteurs d'orientation conceptuelle et de cohérence définis par Zhu *et al.* [114]. D'autre part, le fait de conserver l'ensemble des relations des RTOs sources a l'avantage de garantir que les connaissances qui y sont décrites sont préservées dans l'UMLS mais engendre des cas d'incohérence et d'inexactitude, comme illustrés dans l'étude des relations multiples existant entre une même paire de concepts. Enfin, il est intéressant de souligner qu'en évaluant la qualité de l'UMLS, nous avons pu identifier des erreurs au sein même des RTOs sources, mais aussi formuler des recommandations pouvant être utiles aux développeurs de l'UMLS (*e.g.*, proposer une catégorisation différente pour certains concepts) et des RTOs sources (*e.g.*, ne pas décrire certaines relations lorsqu'elles peuvent être déduites, comme c'est le cas dans la Gene Ontology que nous avons étudiée plus en détails suite à ces constats).

Nous avons vu dans la deuxième section de ce chapitre que le fait de disposer d'une RTO décrite dans un langage formel présente des avantages pour évaluer sa qualité. Le premier travail a montré les possibilités techniques offertes par un langage formel, tandis que l'autre a illustré qu'en disposant d'une RTO dans un tel format, il était possible d'évaluer des connaissances plus riches. Ainsi, nous avons pu réaliser une évaluation de la qualité des relations du NCIt au format RDF en utilisant simplement des requêtes SPARQL. D'autre part, l'évaluation de la GO au format OBO a permis d'identifier que des définitions logiques pour certains de ses concepts. Notons cependant que les approches que nous avons mises en œuvre n'exploitent pas certaines fonctionnalités avancées offertes par les langages formels. En particulier, il est possible de détecter des concepts inconsistants dans une RTO formelle grâce à un raisonneur. Une présentation des travaux existants qui visent à identifier ce type de situations, et parfois les axiomes qui en sont responsables, est disponible dans [150] ansi que dans une revue de la littérature plus récente [151].

Ces quatre études ont démontré la présence d'incohérences dans des RTOs biomédicales. Ces situations problématiques nécessitent d'être identifiées afin de pouvoir être corrigées, ou tout du moins être prises en compte, pour une future utilisation. Le développement de méthodes permettant d'automatiser la détection de potentielles erreurs est indispensable et a fait l'objet de nombreuses propositions. En revanche, l'étape de correction a été moins étudiée [111] car elle nécessite l'intervention d'un humain capable de comprendre le sens des concepts représentés, comme l'ont souligné Schlobach et Cornet [110]. Les travaux de Mortensen *et al.* ont étudié la possibilité de réaliser une partie de cette lourde tâche grâce à la production participative [152] et ont montré qu'elle permettait d'évaluer la qualité de relations de subsomption dans la SNOMED CT aussi bien que des experts [153].

Différentes **perspectives** s'offrent à nous pour aller plus loin dans les recherches sur cette thématique. En particulier, il serait utile de réaliser des évaluations plus exhaustives. En effet, le tableau 2.1 montre que les travaux décrits dans ce chapitre se sont intéressés seulement à certains des facteurs définis par Zhu *et al.* [114]. Même si chaque facteur a été considéré dans au moins une des études présentées, aucune d'entre elles n'a couvert l'ensemble des facteurs. De plus, si l'on reprend les critères définis par Gómez-Pérez pour évaluer la qualité d'une RTO [115], il apparaît que ceux concernant l'évolutivité et la sensibilité n'ont pas été analysés dans le cadre de nos travaux. Ces critères sont particulièrement importants dans un contexte de maintenance des RTO. Il serait d'ailleurs intéressant d'estimer à quel point les méthodes d'évaluation peuvent être utiles pour faciliter cette autre activité du cycle de vie des RTOs, comme mentionné dans notre étude sur la GO ou encore dans [154] sur le NCIt. Des critères supplémentaires ont été recensés par d'autres auteurs [19, 155, 116], comme la clarté (*i.e.*, le sens des termes définis dans la RTO doit être compréhensible), la capacité à raisonner sur une RTO formelle ou encore l'utilité pratique (*i.e.*, le nombre de problèmes concrets auxquels peut répondre la RTO). Cela ouvre un large champ de recherches à poursuivre, avec néanmoins des niveaux de difficulté vraiment différents en fonction des critères à évaluer. Un bon point de départ serait d'exploiter des approches telles que OQuaRE qui est basée sur une norme ISO définissant un processus d'évaluation complet de la qualité des logiciels [156]. Les auteurs ont adapté et enrichi cette norme pour définir un cadre global d'évaluation de la qualité des RTOs via de nombreuses métriques relatives aux diverses caractéristiques des RTOs. L'application d'une telle approche permettrait d'identifier de la manière la plus exhaustive possible les erreurs qu'on peut rencontrer dans une RTO biomédicale donnée.

CHAPITRE 3

INTEROPÉRABILITÉ ENTRE RESSOURCES TERMINO-ONTOLOGIQUES

Les RTOs sont particulièrement nombreuses en santé et il est fréquent de devoir en utiliser plusieurs de concert. Dans ce chapitre, je présente les travaux auxquels j'ai contribué pour garantir l'interopérabilité sémantique entre des RTOs distinctes. Les méthodes proposées ont été appliquées aux domaines de la pharmacovigilance et de la cancérologie.

Pour présenter mes contributions en lien avec cette problématique, je définis tout d'abord les processus permettant une utilisation conjointe de RTOs. En pratique, cela implique nécessairement de trouver des correspondances, souvent désignées par le terme de ***mappings***, entre les entités (les concepts et les relations, principalement) décrites dans des RTOs distinctes. Ces correspondances permettent d'associer ces entités via des relations de subsomption, d'équivalence ou de disjonction [157]. Il existe de nombreuses définitions de ces processus dans la littérature sans réel consensus. Ainsi, les définitions choisies pour qualifier les notions abordées dans ce chapitre sont les suivantes :

— **l'alignement** : ce processus a été défini comme l'identification des mappings existant entre les entités de RTOs distinctes [158]. Ding *et al.* ont ajouté que ces RTOs décrivent des notions communes, impliquant qu'il existe un recouvrement au niveau de leur contenu [159],

— **l'intégration** : le principe consiste à créer une nouvelle RTO à partir des RTOs à intégrer [160] ou à utiliser une RTO de support servant d'intermédiaire afin de trouver des correspondances entre les RTOs à intégrer [161]. L'intégration se fait entre des RTOs représentant des domaines différents mais reliés [111], que la nouvelle RTO ou la RTO de support doivent couvrir. Deux situations peuvent nécessiter ce type de processus [162] : (i) durant la phase de développement où une nouvelle RTO est conçue à partir de RTOs existantes, ou (ii) lorsque des RTOs sont utilisées ensemble pour représenter des données. C'est le deuxième cas de figure qui est considéré dans ce chapitre (une illustration de la première situation est donnée dans le chapitre suivant, en section 4.2).

Une multitude de travaux ont été réalisés sur ces problématiques [111, 158, 160]. En particulier, l'initiative OAEI (Ontology Alignment Evaluation Initiative) [1] organise chaque année depuis

1. `http://oaei.ontologymatching.org/`

2005 des défis pour évaluer des systèmes permettant d'établir des mappings entre RTOs auxquels de nombreux chercheurs participent. Dans ce cadre, Euzenat, Shvaiko et leurs collègues ont réalisé plusieurs revues de la littérature des approches et systèmes existants [157, 163, 164, 165]. Une méta-analyse des différentes revues de la littérature et des travaux passés a été publiée dans le but de faciliter la compréhension de cette thématique très active à des personnes qui y débuteraient leurs recherches [166].

Récemment, Faria *et al.* ont analysé les capacités de certains systèmes à établir des correspondances entre des RTOs biomédicales [167]. En effet, elles font l'objet de tâches spécifiques au sein de l'OAEI depuis 2011, du fait de la complexité du domaine à représenter et des caractéristiques particulières des RTOs biomédicales. Tout d'abord, elles sont très volumineuses, ce qui peut limiter les capacités de raisonnement et rendre ainsi certaines approches d'alignement inutilisables. En revanche, elles disposent d'une sémantique assez simple, comme l'a illustré une étude de Horridge *et al.* en 2011 montrant qu'un bon nombre d'entre elles pouvaient être représentées avec le langage OWL EL [2] [168]. Bien qu'étant moins expressif qu'OWL (*e.g.*, impossibilité d'utiliser la négation, l'union ou encore la symétrie des relations), OWL EL est bien adapté pour représenter les RTOs biomédicales et est supporté par plusieurs raisonneurs [169], tels que ELK. Par ailleurs, la forte composante terminologique des RTOs biomédicales est une opportunité pour les approches lexicales même si la richesse du vocabulaire médical constitue un frein avec les nombreux cas d'ambiguïté, de synonymie ou d'imprécision [13]. Enfin, les RTOs biomédicales peuvent représenter un même domaine de différentes manières. Par exemple, pour coder les diagnostics en cancérologie, il est possible d'utiliser une RTO qui décrit les diagnostics de cancer tels quels ou une RTO qui les représente en combinant la morphologie et la topographie des tumeurs.

Diverses classifications des méthodes existantes ont été proposées dans la littérature, la plus consensuelle étant celle publiée en 2013 par Euzenat et Shvaiko [165], initiée à partir d'une catégorisation d'approches permettant d'aligner des schémas de bases de données [170] puis enrichie pour prendre en compte l'évolution des systèmes qui sont apparus au cours du temps [163, 171]. Cette double classification présente les techniques de base qui existent, en les représentant sous forme de feuilles partagées par deux arbres qui exposent deux points de vue différents (Figure 3.1). La classification descendante distingue les techniques d'après la manière dont celles-ci établissent les correspondances entre RTOs : en considérant les concepts et/ou leurs **instances** (*i.e.*, les données associées aux concepts) de manière isolée (*element-level*) ou selon les relations qu'ils entretiennent avec les autres concepts (*structure-level*). La différenciation entre *sémantique* et *syntaxique* réside dans l'utilisation ou non de ressources formelles ou d'outils exploitant leur sémantique. La classification ascendante sépare les techniques selon qu'elles exploitent des ressources externes (*context-based*) ou qu'elles utilisent uniquement le contenu des RTOs (*content-based*) pour établir les correspondances. Le deuxième niveau pour la catégorie basée sur des ressources externes est le même que pour la classification descendante. En revanche, pour la catégorie *content-based*, les quatre sous-catégories correspondent au type de données traitées par les techniques qui sont qualifiées de : (i) *terminologiques* si ce sont les termes associés aux concepts qui sont exploités, (ii) *structurelles* quand les relations entre les concepts sont prises en compte, (iii) *extensionnelles* si ce sont les instances des concepts qui sont utilisées, et (iv) *sémantiques* lorsque ce sont les connaissances décrites de manière formelle qui sont considérées.

2. `https://www.w3.org/TR/owl2-profiles/#OWL_2_EL_2`

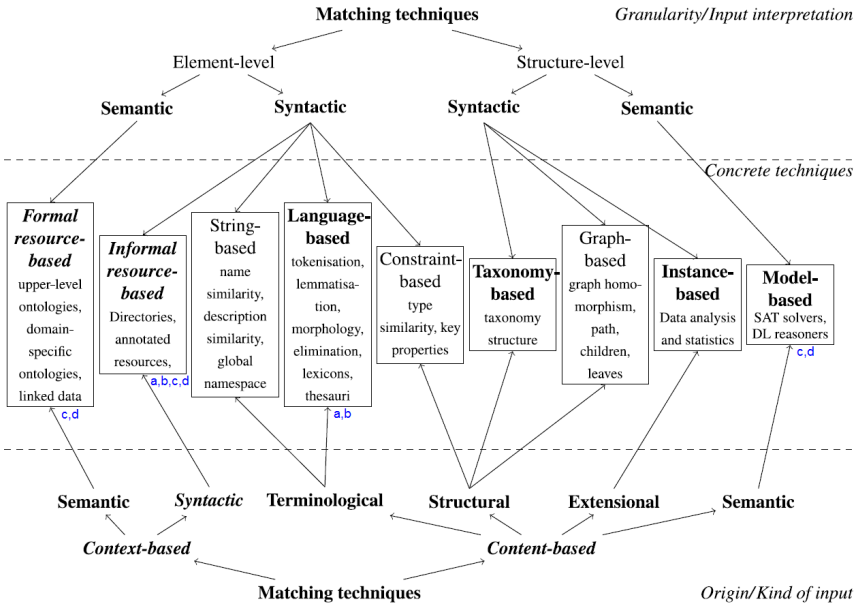

FIGURE 3.1 – Catégorisation des approches d'alignement par Euzenat et Shvaiko (source : [165]). Les catégories n'apparaissant pas en gras sont celles qui avaient été proposées par Rahm et Bernstein en 2001 [170], celles en gras sont issues de la catégorisation définie par Euzenat et Shvaiko en 2007 [163] et celles en gras italique ont été introduites par les mêmes auteurs en 2013 [165].

Ce chapitre expose les différents travaux réalisés en fonction des RTOs qui devaient être utilisées conjointement. Pour chacun d'entre eux, je préciserai le type des techniques que nous avons mises en œuvre d'après la classification d'Euzenat et Shvaiko [165] afin d'expliquer leur positionnement dans la figure 3.1. Je présente dans la section 3.1 deux travaux visant à aligner deux RTOs biomédicales décrivant des connaissances similaires dans le contexte de la pharmacovigilance (a,b). Dans la section 3.2, j'aborde la problématique des RTOs diagnostiques utilisées en cancérologie. Après avoir introduit le contexte (partie 3.2.1) et les RTOs considérées (partie 3.2.2), je décris dans la partie 3.2.3 le modèle ontologique créé par Vianney Jouhet pour leur intégration (c). Enfin, je détaille dans la partie 3.2.4 une des contributions du travail de thèse de Jean Noël Nikiema qui a proposé une solution alternative à celle de Vianney en réutilisant une RTO existante pour intégrer les mêmes RTOs diagnostiques (d). L'étude (e) relève de ces mêmes thématiques mais n'est pas détaillée dans ce manuscrit. Elle présente une revue de la littérature recensant les travaux en lien avec les deux processus d'interopérabilité sémantique abordés dans ce chapitre au regard des conflits sémantiques qu'ils permettent de résoudre.

(a) Fleur Mougin, Marie Dupuch, Natalia Grabar. Improving the mapping between MedDRA and SNOMED CT. Proceedings of the 8th Conference of Artificial Intelligence in Medicine in Europe, Lecture Notes in Artificial Intelligence, 6747 ; pages 220-224, 2011

(b) Fleur Mougin, Natalia Grabar. Using a cross-language approach to acquire new mappings between two biomedical terminologies. Proceedings of the 9th Conference of Artificial Intelligence in Medicine in Europe, Lecture Notes in Artificial Intelligence, 7885 ; pages 221-226, 2013

(c) Vianney Jouhet, Fleur Mougin, Bérénice Brechat, Frantz Thiessard. Building a model for disease classification integration in oncology, an approach based on the National Cancer Institute thesaurus. Journal of Biomedical Semantics, 8(1):6, 2017

(d) Jean Noël Nikiema, Vianney Jouhet, Fleur Mougin. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. Journal of Biomedical Informatics, 74:46 – 58, 2017

(e) *Jean Noël Nikiema, Fleur Mougin, Vianney Jouhet. Finding the appropriate transversal relations between entities of distinct knowledge resources:a step forward on semantic integration. Soumis au Journal of Biomedical Semantics, 2019*

3.1 Alignement de ressources termino-ontologiques

Cette section présente des travaux appliqués au domaine de la pharmacovigilance. Après avoir mentionné succinctement une étude faite lors d'un projet européen du 7ème programme cadre et valorisée par plusieurs publications [74, 75, 76, 77], je décris en détail deux travaux sur l'alignement de RTOs que j'ai réalisés avec Natalia Grabar dans le cadre d'un projet ANR [98, 99].

3.1.1 Contexte médical : la pharmacovigilance

"*La pharmacovigilance est la surveillance des médicaments et la prévention du risque d'effet indésirable résultant de leur utilisation, que ce risque soit potentiel ou avéré*"[3]. En France, elle repose sur la déclaration spontanée des effets indésirables "*par les professionnels de santé, les patients et associations agréées de patients et les industriels avec l'appui du réseau des 31 centres régionaux de pharmacovigilance*". Cette information très précieuse est néanmoins assez pauvre car elle est peu renseignée. Il y a diverses explications à cette sous-déclaration, telles que le temps que cela représente pour les professionnels de santé et la difficulté à identifier la cause réelle des symptômes et les lier au "bon" médicament. Pour pallier ce problème, de nombreux travaux visant à détecter automatiquement les effets indésirables des médicaments ont émergé [172].

C'est dans ce contexte que se positionnait le projet européen EU-ADR. Son objectif était le développement et la validation d'un système informatisé exploitant les données de dossiers de santé électroniques et de bases de données biomédicales pour la détection précoce de vingt-trois effets indésirables de médicaments [173]. Chacune des huit bases de données utilisées possédait des caractéristiques propres dépendant de l'objectif initial pour lequel elle avait été créée et de la fonction locale qu'elle assurait (*e.g.*, médico-économique, soins, audit) et leurs données étaient codées suivant quatre RTOs différentes. L'une de nos tâches consistait à permettre une interrogation harmonisée de ces bases. Comme ces RTOs étaient toutes intégrées dans l'UMLS, nous avons utilisé le metathesaurus afin d'identifier les concepts correspondant aux effets indésirables étudiés dans le cadre du projet [74, 75, 76, 77]. Ensuite, il a suffi de fournir à chaque gestionnaire de bases de données une liste des codes de la RTO appropriée appartenant aux concepts UMLS jugés d'intérêt. Comme des données étaient disponibles en texte libre au sein de certaines bases de données, nous avons également généré une liste de tous les termes synonymes dans la langue du pays concerné qui étaient regroupés dans les concepts UMLS d'intérêt.

La RTO préconisée au niveau européen pour coder les effets indésirables des médicaments est MedDRA [174] mais celle-ci n'est pas utilisée pour coder les diagnostics dans les bases de données hospitalières. Aux États-Unis, c'est la SNOMED CT [175] qui sert pour coder les documents cliniques tandis qu'en France, c'est la SNOMED internationale [176] qui est censée être employée pour coder les documents textuels. Pour détecter automatiquement des effets indésirables, il faut pouvoir comparer ce qui est enregistré dans les déclarations spontanées (où le codage est fait avec MedDRA) et ce qui est retrouvé dans les documents cliniques (codés avec SNOMED CT ou SNOMED internationale) afin de distinguer les effets déjà connus de potentiels effets indésirables des médicaments non encore connus. Ces trois RTOs sont intégrées dans l'UMLS mais nous avons constaté que les mappings entre les concepts de ces RTOs étaient incomplets. Nous avons donc mené deux études visant à enrichir les mappings existant dans l'UMLS [98, 99].

3. `http://ansm.sante.fr/Declarer-un-effet-indesirable/Pharmacovigilance`

3.1.2 Ressources termino-ontologiques considérées

MedDRA. Medical Dictionary for Regulatory Activities (MedDRA) a été conçu pour coder les effets indésirables de médicaments. Il inclut environ 68 000 termes (signes et symptômes, diagnostics, indications thérapeutiques, investigations complémentaires, interventions médicales et chirurgicales, histoire familiale et sociale) structurés suivant cinq niveaux hiérarchiques : les *System Organ Classes*, les *High Level Group Terms*, les *High Level Terms*, les *Preferred Terms* et les *Low Level Terms*.

SNOMED CT. La Systematized NOmenclature of MEDicine - Clinical Terms (SNOMED CT) est une RTO visant à décrire l'ensemble des connaissances biomédicales organisées suivant 19 concepts de haut niveau. Cette RTO contient près de 300 000 concepts qui sont reliés par des relations de subsomption et associatives. Les concepts de la SNOMED CT peuvent être combinés, pour décrire une notion clinique complexe. Ce mécanisme, qualifié de **post-coordination**, garantit une grande expressivité de la SNOMED CT car il permet la description de notions cliniques non encore représentées [177].

SNMI. La Systematized NOmenclature of MEDicine - International (SNMI) est une des RTOs à partir de laquelle a été constituée la SNOMED CT. Elle contient plus de 109 000 concepts organisés de manière hiérarchique et qui peuvent aussi être combinés suivant le mécanisme de post-coordination.

3.1.3 Approche lexicale

Dans le travail réalisé en 2011 avec Natalia Grabar et une de ses doctorantes, nous nous sommes focalisées sur l'alignement entre MedDRA et SNOMED CT [98]. Dans l'UMLS, seulement 42% des concepts de ces deux RTOs étaient reliés. Nous avons proposé une méthode lexicale basée sur la segmentation des termes MedDRA pour améliorer cet alignement (Figure 3.2).

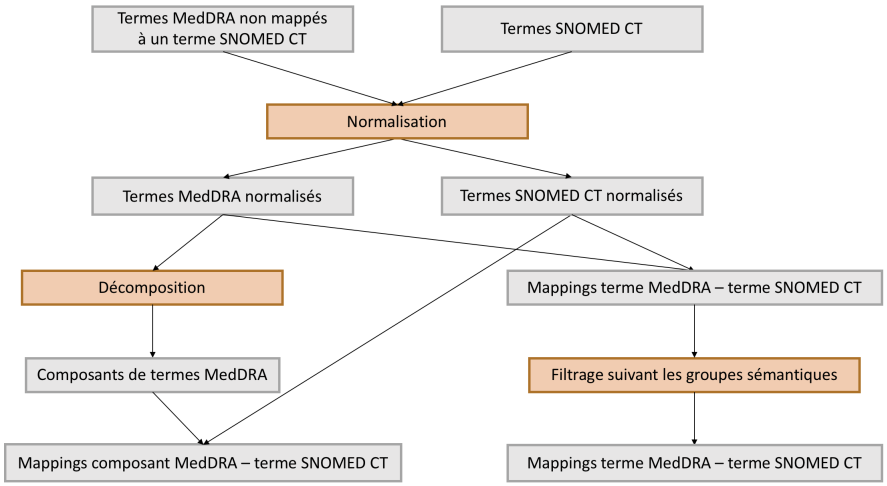

FIGURE 3.2 – Recherche de mappings complets entre chaque composant d'un terme MedDRA et un terme SNOMED CT (à gauche) et de mappings entre un terme MedDRA et un terme SNOMED CT (à droite).

Sélection des termes. Nous avons commencé par identifier les concepts UMLS qui contenaient au moins un terme de MedDRA et aucun terme issu de SNOMED CT. De là, nous avons récupéré tous les termes MedDRA appartenant à ces concepts résultant en une liste de termes MedDRA auxquels aucun terme SNOMED CT n'était associé. Parallèlement, nous avons constitué une liste de l'ensemble des termes SNOMED CT appartenant à au moins un concept UMLS.

Préparation et mapping des termes. L'approche lexicale implémentée consistait en deux étapes. Tout d'abord, les termes ont été segmentés en mots puis normalisés pour prendre en compte l'ordre des mots, la ponctuation, les mots vides, les formes fléchies et dérivées ainsi que les synonymes. Pour cela, nous avons utilisé plusieurs ressources : une liste de mots vides faite par nous-mêmes et celle de la NLM [4], un lexique morphosyntaxique et un lexique de synonymes, constitués tous deux à partir de ressources créées par Natalia Grabar lors de travaux précédents [178, 179] que l'on a enrichies grâce à des synonymes récupérés dans l'UMLS.

Les termes MedDRA normalisés ont tout d'abord été recherchés tels quels puis ils ont été décomposés et normalisés de la même manière que les termes. Trois ensembles de composants issus de la segmentation des termes MedDRA ont été créés (illustrés par le terme *Ear and labyrinth disorders*) :

— ensemble **segmenté** : décomposition suivant les mots vides (deux composants : *ear* et *labyrinth disorders*),

— ensemble **segmenté par coordination** : décomposition suivant les mots vides avec un traitement particulier en présence de coordination (deux composants : *ear disorders* et *labyrinth disorders*),

— ensemble **segmenté syntaxiquement** : décomposition suivant la nature grammaticale et syntaxique des mots (quatre composants : *ear*, *and*, *labyrinth* et *disorders*).

Filtrage des mappings suivant les groupes sémantiques. Comme dit précédemment, les groupes sémantiques ont été définis pour constituer une partition des concepts UMLS. Nous avons exploité cette ressource pour filtrer les mappings obtenus entre un terme MedDRA et un terme SNOMED CT. Ainsi, lorsque le groupe sémantique du concept UMLS auquel appartenait le terme MedDRA était différent de celui contenant le terme SNOMED CT, le mapping a été éliminé automatiquement.

Évaluation des mappings. Dans un premier temps, nous avons compté le nombre de termes MedDRA et le nombre de composants qui avaient pu être mappés à un terme SNOMED CT. Nous avons également calculé le nombre de termes MedDRA dont tous les composants avaient pu être associés à un ou plusieurs termes SNOMED CT, correspondant à des mappings que nous avons qualifiés de "mappings complets". Ensuite, nous avons comparé les mappings complets obtenus par les trois approches de décomposition. Enfin, nous avons examiné manuellement les mappings 1-1 (i.e., associant un terme MedDRA à un terme SNOMED CT). Nous les avons catégorisés comme corrects, liés hiérarchiquement (*i.e.*, si une relation hiérarchique existait entre les notions décrites par les termes MedDRA et SNOMED CT) ou incorrects.

4. `https://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords/`

Résultats de l'alignement entre MedDRA et SNOMED CT. Pour les 30 023 termes MedDRA non mappés à un terme SNOMED CT dans l'UMLS, nous avons trouvé 308 mappings entre un terme MedDRA et un terme SNOMED CT. Notons que le filtrage par les groupes sémantiques avait éliminé préalablement près d'un quart des mappings. Par exemple, le mapping entre *Pleocytosis measurement* (faisant partie du concept UMLS ayant pour CUI C1509130) du groupe sémantique `PROCEDURES` et *Pleocytosis* (C0151857) du groupe sémantique `DISORDERS` a été supprimé. L'évaluation des 308 mappings 1-1 est la suivante :

— 199 mappings corrects (*e.g.*, Acute myringitis without mention of otitis media (C0155459) et Acute myringitis without otitis media (C0395846)),

— 64 mappings impliquaient des termes reliés hiérarchiquement (*e.g.*, Toadstool poisoning (C0858341) et Mushroom poisoning (C0026865), les champignons vénéneux étant un type de champignons),

— 45 mappings incorrects (*e.g.*, Eighth rib fracture (C0920007) et Fracture of eight OR more ribs (C0272566) qui ont été mappés par erreur car les mots *eight* et *eighth* ont tous deux été normalisés en *eight*).

La décomposition a permis de générer un plus grand nombre de mappings (Tableau 3.1).

	Segmenté	Segmenté par coordination	Segmenté syntaxiquement
Nb de composants MedDRA mappés	1236	1130	2273
Nb de termes MedDRA concernés	1123	1159	5077
Nb de termes SNOMED CT mappés à un composant MedDRA	1142	1087	3640
Nb de mappings complets	52	234	361

TABLEAU 3.1 – Résultats de l'alignement MedDRA-SNOMED CT pour chaque ensemble obtenu par décomposition des termes MedDRA.

La comparaison des résultats obtenus par les trois approches de décomposition (Figure 3.3) a montré un faible recouvrement avec seulement 13 mappings trouvés par chacune d'entre elles. Une illustration est le terme MedDRA Suicide of sibling (C0860090) qui a été mappé aux deux termes SNOMED CT Suicide (C0038661) et Sibling (C0037047). Par ailleurs, 86 mappings ont été obtenus par deux des trois approches de décomposition. Par exemple, le mapping de Chronic osteomyelitis involving hand (C0158384) à Chronic osteomyelitis (C0008707) et Hand (C0018563) n'a pas été retrouvé par l'approche par segmentation syntaxique car le mot *involving* n'a pas été ignoré, étant de la catégorie des verbes (alors que c'est un mot qui a été ignoré par les deux autres approches car il faisait partie des mots vides). Enfin, parmi les 135 mappings obtenus seulement par l'approche par coordination, on peut citer Somatoform and factitious disorders (C0851579) qui a été mappé avec succès à Somatoform disorder (C0037650) et Factitious disorder (C0015480).

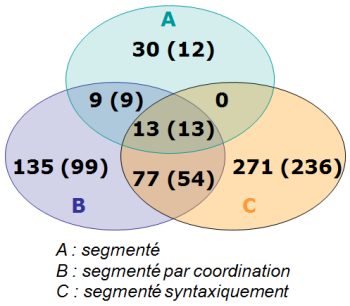

FIGURE 3.3 – Recouvrement des mappings complets obtenus avec les trois approches de segmentation. Les chiffres entre parenthèses correspondent au nombre de ces mappings qui ont été jugés corrects.

Globalement, l'évaluation des mappings complets générés par les différentes décompositions était satisfaisante (Tableau 3.2), indiquant que l'exploitation faite de la compositionnalité des termes MedDRA était utile. Des exemples de mappings corrects étaient : Drain of cerebral subdural space (C0948933) mappé à Drain (C0180499) et Subdural space of brain (C1284568), ou encore Urinary frequency aggravated (C0856128) mappé à Frequency of urination (C0677481) et Aggravated (C0436331). Bien que peu de mappings de termes reliés hiérarchiquement aient été obtenus, ils semblaient intéressants. En effet, le mapping que nous avons trouvé entre Pyrimidine metabolism disorders NEC (C0947944) et Disorder of pyrimidine metabolism (C0268127) nous a permis de remarquer qu'il manquait une relation hiérarchique entre ces deux concepts dans l'UMLS. Les mappings complets incorrects ont révélé deux types d'erreurs récurrentes : lorsque le terme MedDRA contenait une négation et la présence de synonymes incorrects.

	Segmenté		Segmenté par coordination		Segmenté syntaxiquement	
Nb de mappings corrects	34	(65,4%)	175	(74,8%)	303	(84,0%)
Nb de mappings hiérarchiques	12	(23,1%)	43	(18,4%)	33	(9,1%)
Nb de mappings incorrects	6	(11,5%)	16	(6,8%)	25	(6,9%)
Total	52	(100,0%)	234	(100,0%)	361	(100,0%)

Tableau 3.2 – Évaluation qualitative des mappings complets entre MedDRA et SNOMED CT

3.1.4 Approche lexicale et multilingue

Dans le travail de 2013, toujours avec Natalia Grabar, nous avons enrichi la méthode précédente en exploitant les termes de plusieurs langues afin d'améliorer l'alignement entre MedDRA et SNMI [99].

Sélection des termes. Nous avons récupéré la version anglaise de SNMI dans l'UMLS et la version française mise à disposition par l'agence française de la santé numérique (ASIP santé) [5]. Nous avons généré la version espagnole de SNMI à partir de la version espagnole de la SNOMED CT intégrée dans l'UMLS. MedDRA étant disponible dans de nombreuses langues, nous avons exploité les versions française, anglaise et espagnole téléchargeables sur le site Internet de cette RTO [6].

Préparation et mapping des termes. Dans cette étude, nous avons uniquement procédé à l'étape de normalisation des termes français, anglais et espagnols issus de MedDRA et SNMI. Notons qu'il a fallu enrichir les lexiques exploités lors de l'étude précédente pour les langues française et espagnole.

Filtrage des mappings suivant les groupes sémantiques. Ici, nous avons également appliqué ce filtre mais cela a d'abord nécessité d'associer les termes des différentes RTOs avec des concepts UMLS. Pour cela, nous avons exploité les codes puisqu'ils ont l'avantage de ne pas être dépendants de la langue. À cette occasion, nous avons constaté que certains termes n'apparaissaient pas dans l'UMLS, simplement pour des raisons de versions différentes ou parce que des codes sont créés et utilisés uniquement dans certains pays.

Évaluation des mappings. Nous avons comparé les résultats obtenus dans chaque langue et calculé le nombre de mappings communs entre les trois langues (grâce aux codes associés aux termes). Nous avions comme hypothèses que l'usage de différentes langues serait utile dans plusieurs cas : (i) pour enrichir les mappings existant dans une langue donnée, (ii) pour valider les mappings 1-1 s'ils étaient trouvés dans plusieurs langues, (iii) pour désambiguïser les mappings 1-N (si un seul mapping est commun entre les différentes langues, ce mapping 1-1 peut être sélectionné parmi les mappings 1-N trouvés dans certaines langues). Pour l'évaluation qualitative des mappings obtenus, nous avons considéré que des mappings entre deux termes appartenant à un même concept UMLS étaient corrects. Pour les mappings restants, nous avons vérifié si une relation hiérarchique ou de synonymie existait entre les concepts UMLS auxquels appartenaient les termes. Cette évaluation a ainsi pu être réalisée de manière complètement automatique.

Résultats de l'alignement entre MedDRA et SNMI. La méthode basée sur le multilinguisme a permis de générer de nombreux mappings entre MedDRA et SNMI (Tableau 3.3). Le filtrage des mappings d'après les groupes sémantiques a écarté près de 900 mappings automatiquement. Parmi les 9968 mappings restants, 77,7% ont été considérés automatiquement comme

5. `http://esante.gouv.fr/snomed/snomed/`
6. `https://www.meddra.org`

corrects. Une évaluation manuelle aurait été nécessaire concernant les nouveaux mappings (en particulier, ceux qui étaient propres à une seule langue).

	Nb de mappings entre termes de groupes sémantiques différents	Nb de mappings entre termes de concepts UMLS identiques	Nb de nouveaux mappings	Total
Anglais	493	3230	897	4620
Français	250	1506	1063	2819
Espagnol	148	3006	266	3420
Total	891	7742	2226	10859

TABLEAU 3.3 – Nombre de mappings générés entre MedDRA et SNMI dans chaque langue

Indépendamment de la langue, notre approche a permis de trouver 2085 nouveaux mappings entre MedDRA et SNMI. Le recouvrement s'étant avérée assez faible entre les différentes langues (Figure 3.4), il nous a paru pertinent d'exploiter plusieurs langues pour générer un plus grand nombre de mappings. Plus précisément, seulement 6,2% des mappings ont été obtenus par plus d'une langue. Un exemple est le mapping entre les termes MedDRA *Infection due to Mycobacterium fortuitum* (français : *Infection à Mycobacterium fortuitum*, espagnol : *Infección por Mycobacterium fortuitum*) et les termes SNMI *Mycobacterium fortuitum infection* (français : *Infection à Mycobacterium fortuitum*, espagnol : *Infección por mycobacterium fortuitum*), appartenant respectivement aux concepts UMLS C0275711 et C0877567.

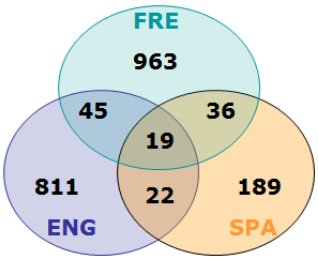

FIGURE 3.4 – Comparaison des nouveaux mappings entre MedDRA et SNMI générés grâce aux termes en français (FRE), en anglais (ENG) et en espagnol (SPA).

Ce faible recouvrement a néanmoins permis de valider 77 mappings 1-1 et de désambiguïser 42 mappings 1-N. Le terme MedDRA *Familial tremor* est une illustration de désambiguïsation car il a été mappé aux termes SNMI suivants en anglais : *Essential tremor*, *Persistent tremor* et *Congenital trembles* alors qu'en français : *Tremblement grossier* (correspondant à *Coarse Tremor* en anglais) et *Tremblement essentiel*. En combinant ces mappings, nous avons sélectionné exclusivement le terme SNMI *Essential tremor*.

3.1.5 Conclusions

Les travaux présentés dans cette partie ont été motivés par le fait que, malgré toute la richesse offerte par l'UMLS, les mappings qui s'y trouvent sont incomplets. Il est donc nécessaire de proposer des méthodes permettant d'enrichir l'alignement entre certaines RTOs sources. Deux approches lexicales différentes ont été décrites : la première a exploité la compositionnalité des termes tandis que la seconde s'est basée sur l'aspect multilingue des RTOs. Par ailleurs, l'utilisation des groupes sémantiques de l'UMLS a été efficace dans les deux travaux pour éliminer automatiquement un nombre non négligeable de mappings erronés. D'après la catégorisation de Euzenat et Shvaiko donnée en figure 3.1, ces études combinent donc des techniques *language-based* (*i.e.*, lexicales) et des techniques *informal resource-based* en exploitant l'UMLS comme ressource externe.

La première approche s'est révélée utile pour identifier plusieurs centaines de nouveaux mappings. Un résultat intéressant est que nous avons obtenu des mappings complets entre un terme MedDRA et plusieurs termes SNOMED CT grâce aux approches de décomposition. Cependant, lors de l'évaluation, nous avons mis en évidence différentes erreurs dues aux ressources lexicales et sémantiques que nous avons utilisées. Une intervention manuelle a donc été nécessaire pour garantir la qualité des mappings créés. La seconde approche a permis de découvrir plus de 2000 nouveaux mappings entre MedDRA et SNMI. L'exploitation de plusieurs langues présente donc un réel intérêt, d'autant plus que le recouvrement entre les différentes langues était faible. La contrepartie est que cela a limité les possibilités de valider ou désambiguïser les résultats obtenus dans une langue grâce à ce qui avait été trouvé dans une autre. Notons que les deux approches proposées auraient pu être combinées pour générer un nombre plus conséquent de mappings candidats, mais cela aurait nécessiter un travail plus lourd de validation manuelle.

Lors de l'utilisation des connaissances présentes dans l'UMLS, nous y avons identifié des problèmes d'inconsistance relevant du facteur d'exactitude défini par Zhu *et al.* [114]. En effet, les deux approches ont montré l'existence de relations de synonymie inappropriées, comme cela a été rapporté dans d'autres études [114]. Une illustration de ce type d'incohérence est la relation de synonymie entre Photophobia aggravated (C0853637) et Photophobia (C0085636) qui devrait en fait être une relation hiérarchique. D'autre part, nous avons observé des situations où des relations de synonymie n'étaient pas définies alors qu'elles devraient l'être, ce qui est aussi un problème connu de l'UMLS [180]. Par exemple, les concepts Morose (C0522168) et Depressed mood (C0344315) devraient être définis comme synonymes (ce que notre méthode est parvenue à établir grâce au multilinguisme). Ici, le critère concerné est donc l'exhaustivité de la couverture, à moins que l'on considère que ces deux concepts ne représentent en réalité qu'une notion, ce qui relèverait alors de l'orientation conceptuelle.

3.2 Intégration de ressources termino-ontologiques

3.2.1 Contexte médical : la cancérologie

En cancérologie, comme dans d'autres domaines biomédicaux, la réutilisation des données est confrontée à la multiplicité et à l'hétérogénéité des RTOs. Les registres des cancers, qui visent à collecter de façon exhaustive les cancers incidents dans la population, sont organisés en réseaux avec une volonté de standardiser ce recueil. C'est la Classification Internationale des Maladies pour l'Oncologie (CIM-O3) qui a été choisie comme RTO commune pour le codage des cancers [181]. Pour assurer leur fonction, les registres recueillent des données provenant de différentes sources codées avec des RTOs multiples. En particulier, la CIM-10 est 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 RTO est utilisée pour l'enregistrement des causes de morbidité et de mortalité. Dans le but d'automatiser l'identification des cas incidents de cancer, il est donc indispensable de pouvoir comparer ce qui est codé en CIM-O3 dans les registres avec ce qui est codé en CIM-10. Cependant, la CIM-O3 décrivant de manière indépendante la morphologie et la topographie des tumeurs, il n'est pas possible d'établir des correspondances exactes (*i.e.*, des équivalences) avec les diagnostics décrits dans la CIM-10, ni même de trouver des liens hiérarchiques entre les concepts de ces deux RTOs.

3.2.2 Ressources termino-ontologiques utilisées

Même si la CIM-10 et la CIM-03 sont qualifiées de classifications des maladies, elles sont utilisées en pratique pour coder un diagnostic. En cancérologie, un diagnostic décrit deux éléments complémentaires de la maladie : le type de cellules tumorales (morphologie) et son site d'apparition (topographie). La CIM-10 et la CIM-O3 permettent toutes les deux de décrire des diagnostics mais nous montrons ci-après que leur structure est différente.

CIM-10. La CIM-10 est la $10^{\text{ème}}$ révision de la classification statistique internationale des maladies et des problèmes de santé connexes organisée suivant 21 chapitres. Le deuxième chapitre correspond aux tumeurs et est divisé en quatre axes disjoints dépendant du comportement de la tumeur : les tumeurs malignes, les tumeurs *in situ*, les tumeurs bénignes et les tumeurs à évolution imprévisible ou inconnue. Parmi les tumeurs malignes, les classes distinguent les tumeurs primitives des tumeurs métastatiques secondaires. La CIM-10 décrit chaque pathologie néoplasique grâce à un seul concept représenté par un code unique. Par exemple, le concept CIM-10 Malignant neoplasm upper-inner quadrant of breast (C50.2) décrit en fait les deux caractéristiques du cancer : 1) son comportement (malin) qui concerne la morphologie de la tumeur et 2) le site d'origine (quadrant supéro-interne du sein) qui correspond à sa topographie.

CIM-O3. La Classification Internationale des Maladies pour l'Oncologie - $3^{\text{ème}}$ révision (CIM-O3) est une RTO qui possède un axe morphologique pour décrire la morphologie des tumeurs et un axe topographique qui précise leur localisation. Un code CIM-O3 de tumeur est composé de 10 caractères correspondant à l'agrégation des quatre caractères du code topographique, suivis des cinq caractères du code morphologique. Les quatre premiers caractères du code morphologique indiquent le type de cellule ou l'histologie, tandis que le cinquième précise le comportement de

la tumeur (un concept morphologique CIM-O3 décrit donc un et un seul comportement). Dans l'absolu, il est possible d'associer n'importe quelle morphologie à n'importe quelle topographie. Pour coder un diagnostic précis de tumeur, la CIM-O3 nécessite de combiner un concept morphologique et un concept topographique. Comme la SNOMED, la CIM-O3 est donc une RTO bénéficiant du mécanisme de post-coordination. Si on reprend l'exemple donné pour la CIM-10, le diagnostic de Malignant neoplasm upper-inner quadrant of breast sera codé en CIM-O3 par la combinaison du concept morphologique Neoplasm, malignant (8000/3) avec le concept topographique Upper-inner quadrant of breast (C50.2).

3.2.3 Intégration par la création d'un modèle

Dans le cadre de la thèse de Vianney Jouhet, nous avons cherché à établir des liens entre des concepts de la CIM-10 et des concepts de la CIM-O3. Nous avons ainsi étudié la faisabilité d'utiliser le NCI thesaurus comme base d'un modèle permettant d'intégrer la CIM-10 et la CIM-O3 et notre étude a montré que cette RTO nécessitait d'être enrichie à cette fin [95]. Ce travail a été initié par le co-encadrement d'un stage de Master 2 qui visait à étudier l'alignement entre la CIM-O3 et le NCI thesaurus et qui a donné lieu à un article présenté aux journées francophones d'informatique médicale en 2014 [182].

3.3.3.1 Ressources externes

NCI thesaurus. Le NCI thesaurus (NCIt) est une RTO de référence au niveau international visant à représenter l'ensemble des connaissances en cancérologie (section 2.2.1). Cette RTO est organisée selon 20 sous-domaines comprenant l'axe `Anatomic structure, system, or substance` qui regroupe les concepts topographiques et l'axe `Neoplasm` où sont décrits les morphologies et les diagnostics. Ainsi, il n'y a pas d'axe spécifique pour décrire les concepts morphologiques et certains diagnostics sont représentés comme des spécialisations anatomiques de morphologies. Cette spécialisation est explicite, à l'image du concept Breast adenocarcinoma qui a pour concepts parents BreastCarcinoma et Adenocarcinoma (qui est une morphologie).

Certains concepts du NCIt sont annotés comme étant mappés à des concepts morphologiques de la CIM-O3 mais il n'est pas précisé s'il s'agit d'un correspondance exacte ou d'un autre type.

Enfin, le NCIt est disponible au format OWL DL [7] qui offre une expressivité maximale des connaissances tout en garantissant la réalisation de l'ensemble des capacités de raisonnement (complétude) en un temps fini (décidabilité).

NCI metathesaurus. Le NCI metathesaurus [8] est une ressource utilisée en cancérologie qui regroupe plus de 75 RTOs biomédicales différentes, dont la CIM-10. Il a été élaboré par le National Cancer Institute (NCI) américain à partir du metathesaurus de l'UMLS, enrichi par d'autres RTOs spécifiques à la cancérologie, comme la CIM-O3. Comme dans l'UMLS, les termes et codes des différentes RTOs représentant un même concept sont enregistrés selon un code unique : le Concept Unique Identifier (CUI).

7. `https://evs.nci.nih.gov/ftp1/NCI_Thesaurus/`
8. `https://ncimeta.nci.nih.gov/ncimbrowser/`

3.3.3.2 Modèle d'intégration basé sur le NCIt

Avant d'intégrer ces RTOs, il a été nécessaire de préciser comment les types de concepts qu'elles représentent devaient être articulés les uns par rapport aux autres. Ensuite, Vianney a conçu un modèle basé sur le NCIt permettant de représenter de manière formelle les concepts de diagnostic, morphologie et topographie ansi que les relations entre ceux-ci (Figure 3.5). Enfin, le modèle a été instancié avec les RTOs à intégrer avant d'évaluer son contenu.

Définition formelle du lien entre diagnostic, morphologie et topographie. La combinaison d'un concept topographique et d'un concept morphologique permettant de décrire un diagnostic, nous avons utilisé les relations suivantes pour spécifier le lien entre ces trois types de concepts :

— *has_morphology* : pour modéliser la relation entre un diagnostic et le type de cellules (morphologie) qui sont impliquées dans la tumeur décrite par le diagnostic,

— *has_primary_site* : pour modéliser la relation entre un diagnostic et un site anatomique (topographie) correspondant au site d'origine de la tumeur décrite par le diagnostic.

Une précision est nécessaire pour la deuxième relation : celle-ci a pour co-domaine un site anatomique particulier, ou une partie de ce site. À titre d'illustration, une tumeur peut apparaître sur le quadrant inféro-externe du sein dans son ensemble, ou bien sur une partie plus précise de ce quadrant. Pour rendre cette description possible, le W3C (World Wide Web Consortium) recommande de créer des concepts anatomiques spécifiques, qualifiés de *Reflexive part*[9], décrivant le concept anatomique et l'ensemble des concepts qui en font partie comme suit :

$$\mathsf{AnatomicConceptReflexivePart} \equiv \mathsf{AnatomicConcept} \sqcup \exists\, part\_of.\mathsf{AnatomicConcept}$$

Grâce à cette modélisation, le concept Malignant neoplasm of the lower-outer quadrant of breast est alors défini comme étant un diagnostic dont la morphologie est une tumeur maligne et dont le site primitif est le quadrant inféro-externe du sein ou une partie de ce quadrant, de la manière suivante :

$$\begin{aligned}&\mathsf{MalignantNeoplasmOfTheLowerOuterQuadrantOfBreast} \equiv\\ &\mathsf{Diagnosis}\\ &\sqcap\ \exists\, has\_morphology.\mathsf{MalignantNeoplasm}\\ &\sqcap\ \exists\, has\_primary\_site.\mathsf{LowerOuterQuadrantOfBreastReflexivePart}\end{aligned}$$

Notons par ailleurs que la CIM-O3 fournit des informations plus fines que la CIM-10. Dans ce cas, il faut établir une correspondance de type hiérarchique entre le concept CIM-O3 et le concept CIM-10 plus générique. Par exemple, l'adénocarcinome du quadrant inféro-externe du sein peut être codé en CIM-O3 mais pas en CIM-10. Dans ce cas, la correspondance avec le concept CIM-10 de tumeur maligne du quadrant inféro-externe du sein a été définie comme suit :

$$\begin{aligned}&\mathsf{Diagnosis}\\ &\sqcap\ \exists\, has\_morphology.\mathsf{Adenocarcinoma}\\ &\sqcap\ \exists\, has\_primary\_site.\mathsf{LowerOuterQuadrantOfBreastReflexivePart}\\ &\sqsubseteq \mathsf{MalignantNeoplasmOfTheLowerOuterQuadrantOfBreast}\end{aligned}$$

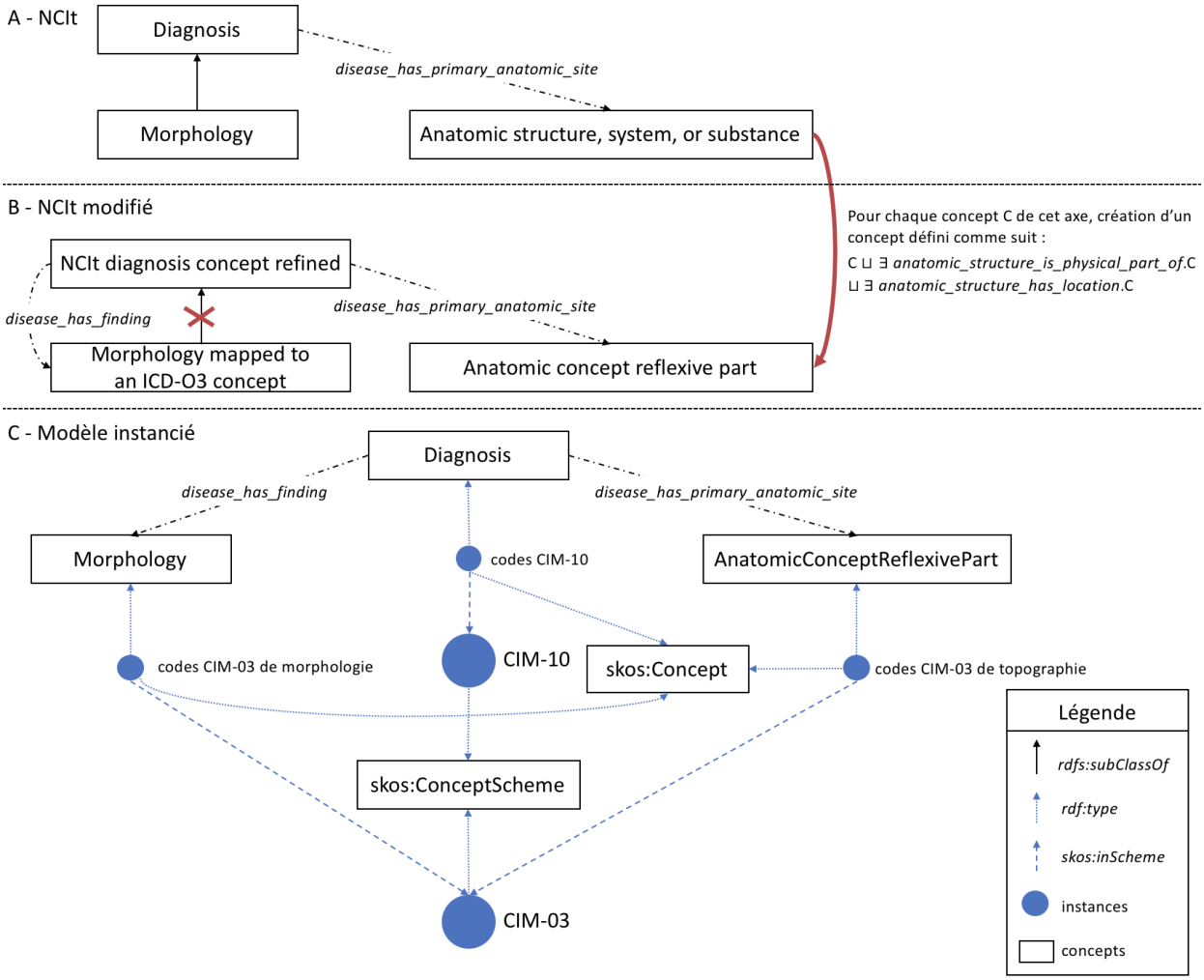

FIGURE 3.5 – Création du modèle en adaptant le NCIt. A présente le NCIt tel qu'il existe, B montre les adaptations effectuées sur le NCIt et C correspond au modèle proposé, instancié par les RTOs CIM-10 et CIM-O3. Celles-ci sont représentées en SKOS et le modèle formel est en OWL.

Construction d'un modèle basé sur le NCIt pour intégrer la CIM-10 et la CIM-O3.

Nous avons pu utiliser la relation du NCIt *disease_has_primary_anatomic_site* telle quelle comme correspondance à la relation *has_primary_site* introduite ci-dessus pour représenter le lien entre un diagnostic et une topographie. En ce qui concerne les concepts topographiques de type *Reflexive part*, nous avons créé un graphe composé de chaque concept de l'axe du NCIt `Anatomic structure, system, or substance` grâce au raisonneur ELK appliqué à tous les concepts de cet axe et en utilisant les deux relations partitives du NCIt, à savoir *anatomic_structure_is_physical_part_of* et *anatomic_structure_has_location*.

En revanche, il n'y avait pas d'axe spécifique pour les morphologies dans le NCIt, ni d'équivalence à la relation *has_morphology*. Cependant, le NCIt fournissait un alignement entre ses diagnostics et les concepts morphologiques de la CIM-O3. Nous avons donc identifié les concepts correspondant à des concepts morphologiques de la CIM-O3 et qui étaient descendants du concept

9. `http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole/`

NCIt Finding (C3367) à partir des mappings du NCIt et nous avons raffiné les concepts de diagnostics grâce à la relation du NCIt *disease_has_finding* de la manière suivante :

NCItDiagnosisConceptRefined ≡
NCItConcept ⊓ ∃ *disease_has_finding*.MorphologyMappedToAnICD-O3Concept

Chacune de ces morphologies a été classée d'après son comportement tumoral (*i.e.*, bénin, malin primaire, in situ, malin métastatique, indéterminé entre bénin ou malin, indéterminé entre primaire ou métastatique).

Grâce à ces enrichissements du NCIt, la définition logique permettant de relier les trois concepts d'intérêt a finalement été la suivante :

Diagnosis ≡
∃ *disease_has_finding*.Morphology
⊓ ∃ *disease_has_primary_anatomic_site*.AnatomicConceptReflexivePart

Le modèle correspondant est présenté en figure 3.5C. Les langages ont été choisis d'après les suggestions de Tao *et al.* [183], à savoir SKOS (Simple Knowledge Organization System) [10] pour la description des RTOs à intégrer et OWL pour représenter les concepts génériques ainsi que les relations entre ceux-ci. Par ailleurs, l'articulation entre SKOS et OWL suit les recommandations du W3C [11].

Pour implémenter ce modèle, nous avons donc construit une expression combinant chaque topographie à chaque morphologie comme suit :

∃ *disease_has_finding*.Morphology
⊓ ∃ *disease_has_primary_anatomic_site*.AnatomicConceptReflexivePart

Si au moins un sous-cconcept de cette expression était identifié par le raisonneur dans le NCIt, le diagnostic équivalent à l'expression a été ajouté au modèle.

Instanciation du modèle. Pour finaliser le modèle, nous l'avons instancié avec les concepts de la CIM-10 et de la CIM-O3. Cette étude s'est focalisée sur la description des tumeurs primaires, ignorant donc les métastases et les tumeurs à comportement incertain.

Pour instancier les diagnostics du modèle, le processus a été le suivant :

1. identification des mappings entre des codes CIM-10 et des concepts NCIt ayant le même CUI dans le NCI metathesaurus,
2. définition des concepts correspondant à ces codes CIM-10 en spécifiant explicitement la restriction sur le comportement tumoral qui est sous-jacent dans la CIM-10 (mais connu grâce à la hiérarchie) et en ajoutant une restriction précisant le site primitif. Par exemple :
   - Breast, Unspecified (C50.9) est une tumeur maligne d'après sa classification dans la CIM-10,
   - Breast, Unspecified (C50.9) a le même CUI que le concept NCIt Malignant Breast Neoplasm (C9335) dans le NCI metathesaurus,
   - Malignant Breast Neoplasm (C9335) a pour site primitif Breast (C12971) dans le NCIt,

10. https://www.w3.org/TR/skos-reference/
11. https://www.w3.org/TR/skos-primer/#secskosowl

— L'expression correspondant à Breast, Unspecified (C50.9) a donc été la suivante :

MalignantBreastNeoplasm $\sqcap \exists$ *disease_has_finding*.MalignantPrimaryNeoplasm $\sqcap \exists$ *disease_has_primary_anatomic_site*.Breast

3. instanciation des diagnostics correspondants dans le modèle grâce au raisonneur.

Pour instancier les topographies de type *Reflexive part* avec la CIM-O3, le principe a été le suivant :

1. identification des mappings entre des codes topographiques CIM-O3 et des concepts NCIt ayant le même CUI dans le NCI metathesaurus,
2. instanciation des topographies de type *Reflexive part* correspondant à ces codes dans le modèle après raisonnement.

Les morphologies ont, elles, été instanciées directement grâce aux mappings fournis par le NCIt.

Évaluation du modèle. Le NCI fournit un ensemble d'outils de conversion entre les différentes CIM. Pour évaluer le modèle proposé, nous avons utilisé un fichier décrivant des mappings entre les concepts CIM-10 et CIM-O3 [12]. En pratique, nous avons reconstruit des combinaisons de morphologie-topographie CIM-O3 mappées à des codes CIM-10 dans ce fichier pour constituer un *gold standard*. Nous avons alors cherché si au moins un concept de diagnostics du modèle était instanciée à la fois par le code CIM-10 et par la combinaison CIM-O3 impliqués dans chacun des mappings du *gold standard*.

3.3.3.3 Résultats de l'intégration de la CIM-10 et la CIM-O3 via le modèle

Modèle construit à partir du NCIt. Un total de 6720 topographies ont été identifiées dans le NCIt et les topographies de type *Reflexive part* correspondantes ont été introduites dans le modèle. En ce qui concerne les morphologies, 1120 codes NCIt étaient reliés aux 1094 codes morphologiques de la CIM-O3. Ces 1094 morphologies ont été ajoutées au modèle et classées automatiquement dans une des six classes morphologiques correspondant aux différents comportements tumoraux. En combinant ces 1100 morphologies aux 6720 topographies, 7 392 000 expressions ont été générées. Parmi celles-ci, 20 133 (0,27%) subsumaient au moins un concept NCIt et ont donc été incluses dans le modèle.

Instanciation du modèle. Le tableau 3.4 montre que seulement une partie des codes des RTOs à intégrer ont permis d'instancier le modèle.

Caractéristiques du modèle final. Le modèle créé contient 113 643 axiomes, incluant 27 953 concepts (6720 topographies, 1100 morphologies et 20 133 diagnostics). Issus des RTOs à intégrer, 1440 codes (278 codes topographiques CIM-O3, 860 codes morphologiques CIM-O3 et 302 codes de diagnostics CIM-10) ont permis d'instancier au moins un de ces concepts. Une grande partie des codes CIM-10 (51%) et des codes topographiques de la CIM-O3 (28%) instanciaient de multiples concepts. Cette situation a été observée lorsque la hiérarchie des diagnostics dans le

12. `https://seer.cancer.gov/tools/conversion/ICD03toICD9CM-ICD10-ICD10CM.xls`

	Nb total de codes	Nb de codes instanciant le modèle
Topographies CIM-O3	409	278 (68,0%)
Morphologies CIM-O3	873	860 (98,5%)
Tumeurs CIM-10	727	302 (41,5%)
Tumeurs malignes CIM-10	481	207 (43,0%)
Tumeurs bénignes CIM-10	180	73 (40,5%)
Tumeurs *in situ* CIM-10	66	22 (33,3%)

TABLEAU 3.4 – Nombre total de codes CIM-10 et de codes CIM-O3 et nombre d'entre eux qui ont intégrés dans le modèle. Les codes CIM-10 des catégories C76-C80 Malignant neoplasms of ill-defined, secondary and unspecified sites et D37-D48 Neoplasms of uncertain or unkown behavior ont été exclus, tout comme les codes CIM-O3 se terminant par /6 qui correspondent aux Malignant neoplasms, stated or presumed to be secondary et par /1 aux Neoplasms of uncertain and unknown behavior.

NCIt n'était pas en accord avec la morphologie et la topographie que nous avions utilisée pour les décrire. Par exemple, Colon Cavernous Hemangioma a été instancié par plusieurs codes car il apparaissait dans le modèle comme un sous-concept direct des expressions suivantes :

$\exists\, disease_has_finding.$CavernousHemangioma
$\sqcap\ \exists\, disease_has_primary_anatomic_site.$ColorectalRegionReflexivePart

$\exists\, disease_has_finding.$CavernousHemangioma
$\sqcap\ \exists\, disease_has_primary_anatomic_site.$ColonReflexivePart

$\exists\, disease_has_finding.$HemangiomaNOS
$\sqcap\ \exists\, disease_has_primary_anatomic_site.$ColorectalRegionReflexivePart

$\exists\, disease_has_finding.$HemangiomaNOS
$\sqcap\ \exists\, disease_has_primary_anatomic_site.$ColonReflexivePart

L'explication de ce type de situation est double : 1) Colon Cavernous Hemangioma était décrit comme ayant Colon et Colorectal Region comme sites primitifs, et 2) il n'y avait pas de relation *anatomic_structure_has_location* ni de relation *anatomic_structure_is_physical_part_of* entre ces deux concepts topographiques. Par ailleurs, ce diagnostic était décrit dans le NCIt comme un concept enfant de Cavernous hemangioma et Hemangioma, NOS.

Comparaison avec le *gold standard*. Construit à partir du fichier fourni par le NCI, le *gold standard* comportait 157 550 mappings entre un code CIM-10 et une combinaison topographie-morphologie de la CIM-O3. Comme il n'a pas été possible d'instancier tous les concepts du modèle en l'absence de certains mappings de codes CIM-10 et CIM-O3 avec un code NCIt dans le NCI metathesaurus, 59% des mappings du *gold standard* n'ont pas pu être évalués. Le modèle a permis de relier 42 260 mappings (soit 65% des mappings restants) à au moins un diagnostic. Une part importante de ces mappings (36%) étaient reliés à plusieurs concepts de diagnostics.

3.2.4 Intégration via une ressource termino-ontologique existante

Le travail présenté dans cette section s'est fait dans la continuité de l'étude précédente ; le contexte et les RTOs à utiliser conjointement étaient les mêmes mais la RTO considérée pour établir les correspondances était la SNOMED CT. Contrairement à l'approche décrite dans la partie précédente, le but du travail de Jean Noël était d'utiliser la SNOMED CT telle quelle pour intégrer la CIM-10 et la CIM-O3 [184, 185, 57].

3.2.4.1 Une ressource termino-ontologique comme support à l'intégration

J'introduis d'abord le cadre conceptuel utilisé pour réaliser l'intégration de la CIM-10 et de la CIM-O3. Je décris ensuite la RTO que nous avons choisie comme support à ce processus puis je détaille la mise en œuvre des phases d'ancrage et de dérivation.

Sélection des concepts CIM-10 et CIM-O3. Préalablement à l'intégration, la liste exhaustive des codes CIM-10 (codes C00 à D48) et des codes CIM-O3 a été collectée à partir du NCI metathesaurus. Les codes d'en-tête (*e.g.*, C00-C97 Malignant neoplasms) qui ne sont pas utilisés en pratique pour coder des diagnostics ont été retirés de cette liste.

Cadre conceptuel. Le cadre conceptuel choisi se base sur une RTO, qualifiée de **support**, et a été défini pour faciliter l'alignement de RTOs dont la structure est limitée à de simples hiérarchies [161, 186]. Cette RTO de support doit non seulement contenir les connaissances nécessaires pour couvrir les domaines des RTOs à aligner mais aussi avoir une structure plus riche que ces RTOs. Ce cadre, que nous avons utilisé dans le but de faire l'intégration entre les RTOs CIM-10 et CIM-O3 comme illustré dans la figure 3.6, comporte les deux étapes suivantes :

— la phase d'ancrage qui vise à retrouver des mappings candidats, nommés **ancres**, entre les concepts des RTOs à intégrer et la RTO qui sert de support,

— la phase de dérivation qui consiste à identifier les relations existant entre les concepts participant aux ancres au sein de la RTO de support afin d'identifier des correspondances entre les concepts des RTOs à intégrer.

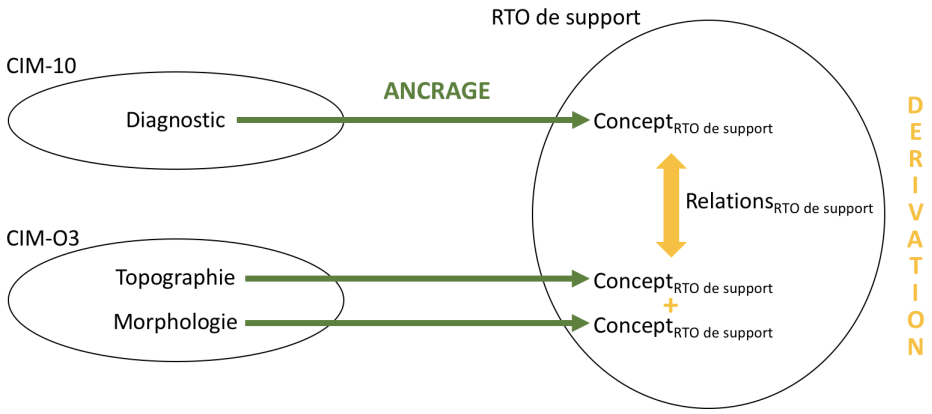

FIGURE 3.6 – Cadre conceptuel pour l'intégration des RTOs CIM-10 et CIM-O3 via une RTO de support consistant en deux phases : l'ancrage et la dérivation.

Choix de la RTO de support. Nous avons choisi la SNOMED CT comme RTO de support pour deux raisons : sa couverture du domaine et ses caractéristiques ontologiques. Cette RTO vise en effet à décrire de manière exhaustive l'ensemble des connaissances du champ de la santé. Comme dit précédemment, elle est organisée suivant 19 concepts de haut niveau, incluant Clinical Finding dont Disease est l'un des descendants, et Body structure qui a parmi ses descendants les concepts Proliferative mass décrivant des concepts morphologiques et Anatomical structure décrivant des concepts topographiques. La SNOMED CT permet donc *a priori* de représenter à la fois les concepts de la CIM-10 et ceux de la CIM-O3. Par ailleurs, la SNOMED CT associe des définitions logiques à la plupart de ses concepts. Il est ainsi possible de décrire une tumeur dans la SNOMED CT grâce aux liens sémantiques suivants :

— *associated_morphology* qui a pour domaine une maladie et pour co-domaine une lésion histologique (*i.e.*, une morphologie),

— *finding_site* qui a pour domaine une maladie et pour co-domaine une localisation anatomique (*i.e.*, une topographie).

Le choix de la SNOMED CT a également été motivé par le fait qu'elle fournit des tables de mapping (nommées *SNCTmt* dans la suite de cette partie) entre ses concepts et ceux de la CIM-10 mais aussi ceux de la CIM-O3. Ces mappings ont été établis manuellement avec pour objectif d'associer à chaque code SNOMED CT un ou plusieurs code(s) CIM-10 ou CIM-O3.

Phase d'ancrage. Cette phase a consisté en trois étapes (Figure 3.7) : l'identification des mappings candidats, le filtrage des ancres et la désambiguïsation des ancres multiples.

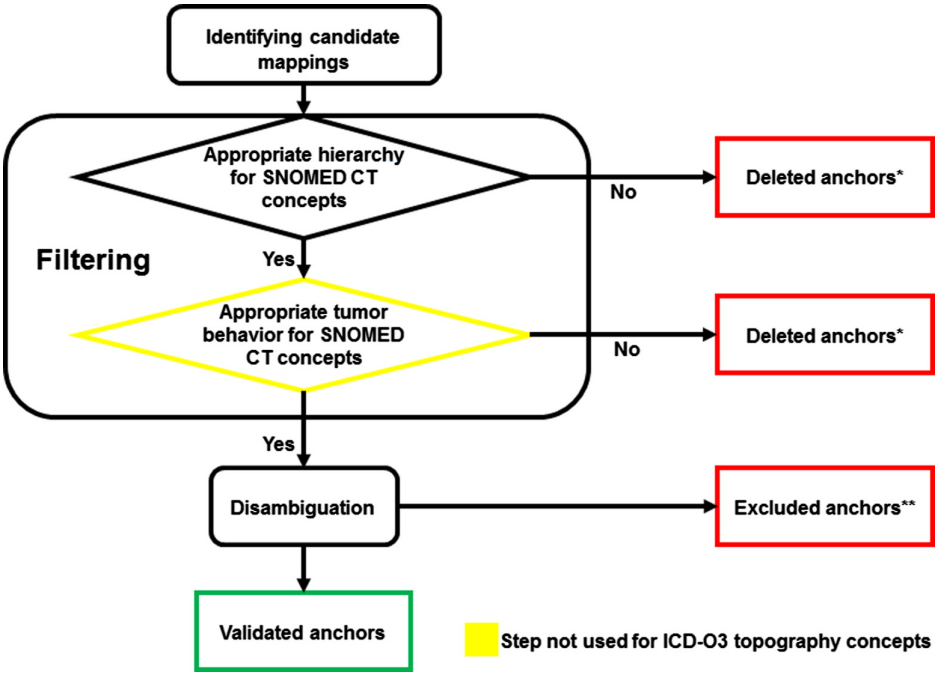

FIGURE 3.7 – Les trois étapes de la phase d'ancrage (source : [57]) : l'identification des mappings candidats, le filtrage des ancres et la désambiguïsation des ancres multiples. Les ancres qui ont été supprimées (*) correspondaient à des mappings erronés tandis que les ancres exclues (**) étaient des mappings corrects mais ne dénotant pas une équivalence entre les concepts mappés.

Deux ressources ont été utilisées pour la sélection des mappings candidats : les tables de mapping SNCTmt et le NCI metathesaurus afin d'identifier les CUI incluant à la fois un code SNOMED CT et un code CIM-10 ou CIM-O3. Ces mappings candidats constituaient les ancres.

Pour éliminer les ancres incorrectes, nous avons effectué deux types de **filtrage** : d'après la hiérarchie de la SNOMED CT et d'après le comportement tumoral. Le filtrage selon la hiérarchie visait à supprimer les ancres impliquant des concepts qui ne représentaient pas les mêmes notions cliniques. Ainsi, le mapping était considéré erroné dans les cas suivants :

— pour les concepts CIM-10 : si le concept SNOMED CT n'était pas un descendant du concept Disease (64572001),

— pour les concepts CIM-O3 de morphologie : si le concept SNOMED CT n'était pas un descendant du concept Proliferative mass (416939005),

— pour les concepts CIM-O3 de topographie : si le concept SNOMED CT n'était pas un descendant de Anatomical structure (91723000).

Le filtrage selon le comportement tumoral a été appliqué uniquement aux ancres impliquant un concept CIM-10 ou un concept CIM-O3 de morphologie. Ce filtrage consistait à supprimer toutes les ancres impliquant des concepts ne décrivant pas le même comportement tumoral. Pour cela, Jean Noël a identifié manuellement les concepts SNOMED CT correspondant aux différentes classes de comportements tumoraux qui sont représentées dans la CIM-10 et dans l'axe morphologique de la CIM-O3 (Tableau 3.5).

	Classes de comportement tumoral	Concepts SNOMED CT correspondants
CIM-10	Primary malignant (C00-C75)	Primary malignant neoplasm (372087000)
	Secondary malignant (C76-C80)	Secondary malignant neoplastic disease (128462008) Malignant tumor of unknown origin or ill-defined site (302817000)
	Haematological malignancy (C81-C96)	Malignant tumor of lymphoid, hemopoietic AND/OR related tissue (269475001)
	Multiple tumors (C97)	Multiple malignancy (363500001)
	Tumor in situ (D00-D09)	Carcinoma in situ (109355002) Melanoma in situ by body site (127330008)
	Benign tumor (D10-D36)	Benign neoplastic disease (20376005)
	Unpredictable tumor (D37-D48)	Neoplastic disease of uncertain behavior (118616009)
CIM-03	Benign (/0)	Neoplasm, benign (3898006)
	Undetermined behavior (/1)	Neoplasm, uncertain whether benign or malignant (86251006)
	Uncertain or unknown tumor behavior (/9)	Neoplasm, malignant, uncertain whether primary or metastatic (6219000)
	In situ morphology (/2)	In situ neoplasm (127569003)
	Primary malignant morphology (/3)	Malignant neoplasm, primary (86049000)
	Secondary malignant morphology (/6)	Neoplasm, metastatic (14799000)

Tableau 3.5 – Concepts SNOMED CT correspondant aux classes de comportement tumoral de la CIM-10 et de la CIM-O3.

Le processus de **désambiguïsation** avait pour but de choisir une seule ancre pour les cas où un concept CIM-10 ou CIM-O3 était mappé à plusieurs concepts SNOMED CT. S'il existait une relation de subsomption entre les concepts SNOMED CT concernés, seule l'ancre impliquant le concept SNOMED CT le plus général a été gardée. Ce processus a d'abord été appliqué aux ancres issues des SNCTmt et du NCI metathesaurus, indépendamment. Dans un deuxième temps, les ancres désambiguïsées obtenues via chacune de ces ressources ont été regroupées et une deuxième désambiguïsation a été réalisée lorsque c'était nécessaire.

Pour évaluer les méthodes implémentées lors de cette phase, nous avons d'abord étudié la couverture des concepts CIM-10 et CIM-O3 au sein des ancres et comparé les résultats obtenus via les SNCTmt et le NCI metathesaurus. Nous avons ensuite mesuré l'impact des différentes étapes de la phase d'ancrage en déterminant le nombre d'ancres obtenues pour chaque concept CIM-10 ou CIM-O3 suivant leur cardinalité :

— ancres 1-1 : un concept CIM-10 ou CIM-O3 a été mappé à un concept SNOMED CT,

— ancres 1-N : un concept CIM-10 ou CIM-O3 a été mappé à plusieurs concepts SNOMED CT,

— ancres 1-0 : un concept CIM-10 ou CIM-O3 n'a pu être mappé à aucun concept SNOMED CT.

Phase de dérivation. Seules les ancres de cardinalité 1-1 ont été utilisées dans la phase de dérivation. Ainsi, chaque combinaison possible d'un concept CIM-O3 de morphologie et d'un concept CIM-O3 de topographie correspondait à une combinaison unique de deux concepts SNOMED CT. Pour chacune de ces combinaisons, Jean Noël a cherché un concept SNOMED CT descendant du concept Finding et qui était équivalent à ou, à défaut, parent de chaque élément de la combinaison via les relations appropriées (Figure 3.8). Pour cela, il a exploité la version de la SNOMED CT disponible au format OWL DL et généré automatiquement des requêtes (*i.e.*, *DL-queries*) exécutées sur la structure inférée de la SNOMED CT. Le raisonneur choisi était ELK étant donné sa capacité à classer cette RTO efficacement et rapidement [187]. Jean Noël a finalement vérifié si les concepts SNOMED CT obtenus étaient ancrés à un concept CIM-10.

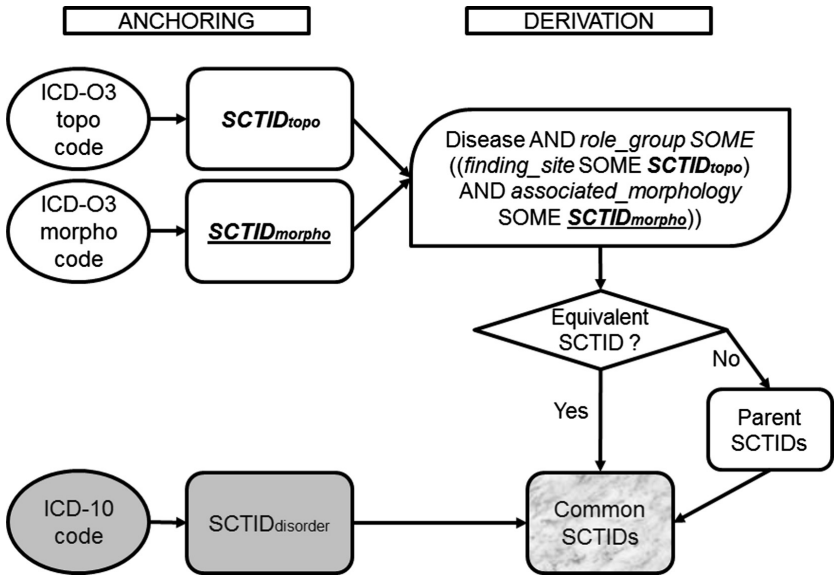

FIGURE 3.8 – La phase de dérivation (source : [57]) : identification des concepts SNOMED CT de maladie permettant d'établir des correspondances entre les concepts CIM-10 et CIM-O3 (SCTID correspond à l'identifiant des concepts SNOMED CT).

Pour l'évaluation de cette phase, une analyse quantitative et une analyse qualitative des résultats obtenus ont été réalisées. Pour l'analyse quantitative, en plus de la couverture des concepts CIM-10 et CIM-O3 impliqués dans la dérivation, le nombre de dérivations trouvées pour chaque concept CIM-10 a été calculé d'après les cardinalités suivantes :

— dérivations 1-1 : un concept CIM-10 était associé à une combinaison unique de concepts CIM-O3 de morphologie et de topographie,

— dérivations 1-N : un concept CIM-10 était associé à plusieurs combinaisons de concepts CIM-O3 de morphologie et de topographie,

— dérivations 1-0 : un concept CIM-10 n'avait pu être associé à aucune combinaison de concepts CIM-O3 de morphologie et de topographie.

Pour l'analyse qualitative, les résultats ont été comparés avec le même fichier du NCI que dans le travail précédent. Nous avons ainsi déterminé le recouvrement de nos résultats avec les 23 694 combinaisons de concepts CIM-O3 de morphologie et de topographie du *gold standard*.

3.2.4.2 Résultats de l'intégration de la CIM-10 et de la CIM-O3 via la SNOMED CT

Au total, 852 concepts CIM-10, 330 concepts CIM-O3 de topographie et 1032 concepts CIM-O3 de morphologie ont été traités dans cette étude.

Phase d'ancrage. La figure 3.9 présente le nombre de concepts CIM-10 et CIM-O3 impliqués dans des ancres. En considérant les deux ressources utilisées pour établir ces ancres, plus de 88% des concepts CIM-10 et CIM-O3 ont été mappés à un concept SNOMED CT, avec un pourcentage ayant atteint 99,3% pour les concepts CIM-O3 de morphologie.

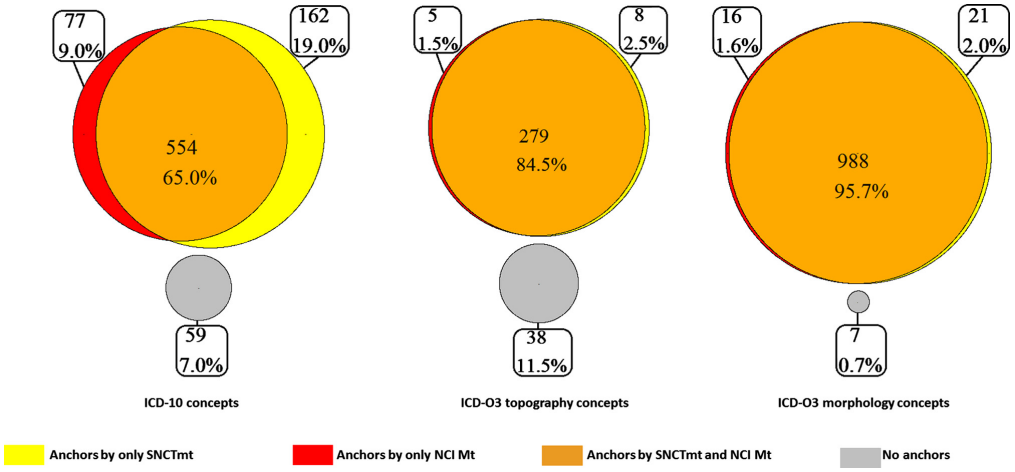

FIGURE 3.9 – Nombre et pourcentage de concepts CIM-10 et CIM-O3 impliqués dans des ancres, selon la ressource utilisée pour établir ces ancres (source : [57]). La taille des cercles est proportionnelle au pourcentage de recouvrement des RTOs (NCI Mt désigne le NCI metathesaurus).

Le tableau 3.6 présente les résultats obtenus lors de l'étape de filtrage. Le filtrage selon la hiérarchie ne s'est pas avéré très utile pour ce qui est des ancres obtenues par les SNCTmt. En revanche, cela a été efficace pour filtrer les ancres issues du NCI metathesaurus. En ce qui concerne le filtrage suivant le comportement tumoral, il a eu globalement pour effet de diminuer le nombre de concepts impliqués dans des ancres de cardinalités 1-1 et 1-N, excepté pour les concepts CIM-10 dont la participation dans les ancres (obtenues via les SNCTmt) de cardinalité 1-1 a augmenté.

L'impact de l'étape de désambiguïsation est détaillé dans le tableau 3.7. Après regroupement des ancres issues des ressources utilisées pour les établir, le nombre de concepts CIM-10 et CIM-O3 impliqués dans des ancres a augmenté. Au final, le nombre de concepts utilisés dans la phase suivante (car impliqués dans des ancres de cardinalité 1-1) était donc de : 487 concepts CIM-10 (57,2%), 127 concepts CIM-O3 de topographie (38,5%) et 901 concepts CIM-O3 de morphologie (87,3%). Notons que la couverture des concepts CIM-10 et celle des topographies CIM-O3 sont corrélées car les diagnostics de cancer dans la CIM-10 sont regroupés suivant la localisation anatomique des tumeurs. Ainsi, l'absence d'ancres pour un concept CIM-O3 de topographie

Étapes	Cardinalités des ancres	CIM-10		CIM-O3			
				Topographie		Morphologie	
		SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Initial	1-1	79	516	4	132	960	539
	1-N	637	115	283	152	49	465
	1-0	136	221	43	46	23	28
Filtrage selon la hiérarchie	1-1	79	572	4	125	959	847
	1-N	637	48	282	130	49	150
	1-0	136	232	44	75	24	35
Filtrage selon le comportement tumoral	1-1	159	288			912	838
	1-N	507	27			48	103
	1-0	186	537			72	91

TABLEAU 3.6 – Distribution des concepts CIM-10 et CIM-O3 impliqués dans des ancres obtenues via les SNCTmt et le NCI metathesaurus (NCI Mt) après chaque étape de filtrage.

a automatiquement résulté en l'absence d'ancres pour les concepts CIM-10 impliquant cette localisation anatomique. En revanche, la large couverture des concepts CIM-O3 de morphologie s'explique par deux aspects : (i) la même lésion histologique peut exister pour des localisations anatomiques différentes, et (ii) la description de ces lésions est plus précise dans la CIM-O3 que dans la CIM-10 (ce qui explique aussi les nombreuses dérivations 1-N).

		Cardinalités des ancres	CIM-10		CIM-O3			
					Topographie		Morphologie	
			SNCTmt	NCI Mt	SNCTmt	NCI Mt	SNCTmt	NCI Mt
Étapes	Avant désambiguïsation	1-1	159	288	4	125	912	838
		1-N	507	27	282	130	48	103
	Après désambiguïsation	1-1	448	302	131	184	957	879
		1-N	218	13	155	71	3	62

TABLEAU 3.7 – Désambiguïsation des ancres obtenues via les SNCTmt et le NCI metathesaurus (NCI Mt), indépendamment. Les ancres 1-0 ne sont pas présentées car leur nombre restait inchangé lors de cette étape.

Phase de dérivation. Parmi les concepts impliqués dans les ancres de cardinalité 1-1, 203 concepts CIM-10 (41,6%) ont pu être associés à 127 concepts CIM-O3 de topographie (100%) et 892 concepts CIM-O3 de morphologie (99%) lors de la phase de dérivation. Sur les 203 dérivations obtenues, quasiment la totalité d'entre elles (192) étaient de cardinalité 1-N. Un exemple des 11 dérivations 1-1 obtenues est le concept CIM-10 Benign neoplasm of duodenum (D13.2) associé à la combinaison du concept CIM-O3 de morphologie Lipoma, NOS (8850/0) avec le concept CIM-O3 de topographie Duodenum (C17.0).

Sur les 157 550 mappings du *gold standard*, 84,8% des mappings n'ont pas pu être évalués en l'absence d'ancrages des concepts CIM-10 et CIM-O3 impliqués dans les mappings. Notre approche a néanmoins permis de retrouver 11 932 mappings identiques à ceux du SEER (soit 47,8% des mappings restants).

3.2.5 Conclusions

Les techniques mises en œuvre dans les deux travaux de cette section pour permettre l'intégration de deux RTOs biomédicales sont diverses (Figure 3.1) : elles exploitent les deux types de ressources externes et sont donc *formal resource-based* (utilisation du NCIt et de la SNOMED CT décrits en OWL DL et du modèle créé) et *informal resource-based* (mappings fournis par le NCI metathesaurus et les SNCTmt) et sont aussi du type *model-based* puisqu'un raisonneur (ELK) a été nécessaire pour établir les correspondances.

Le premier travail a montré qu'il était possible de créer un modèle à partir du NCIt pour fournir une vue sémantiquement intégrée et exploitable par les machines de deux RTOs qui

ont vocation à être utilisées conjointement. Le NCIt n'a pu être utilisé tel quel car il ne fait pas la distinction entre les morphologies et les diagnostics mais aussi parce que les diagnostics qu'il décrit ont un site primitif mais il n'est pas spécifié que le site anatomique concerné peut également être une partie de ce site (en plus de lui-même). La deuxième étude a illustré que le cadre conceptuel défini initialement pour l'alignement de RTOs est adapté pour mettre en correspondance des RTOs décrivant des notions disjointes et complémentaires. Les méthodes implémentées pour filtrer et désambiguïser les ancres obtenues à partir de ressources externes se sont appuyées sur la structure des RTOs à intégrer qui étaient pourtant peu structurées. Cela a résolu certaines limites des approches utilisées par les ressources externes pour constituer leurs mappings, à savoir une approche morphosyntaxique pour le NCI metathesaurus et manuelle pour les SNCTmt. Les résultats obtenus dans ces deux études sont prometteurs car des mappings complexes impliquant un concept CIM-10 et deux concepts CIM-O3 ont été trouvés. Cependant, l'étape visant à établir les mappings entre les concepts des RTOs à intégrer et les concepts de la RTO utilisée de base au modèle / comme support mérite d'être améliorée. Le travail de Jean Noël couvrait un peu plus de concepts CIM-O3 de morphologie et moins de concepts CIM-10 et CIM-O3 de topographie. Il serait intéressant de combiner les résultats des deux travaux afin d'obtenir une couverture supérieure des RTOs à intégrer. Notons finalement que l'étude réalisée par Vianney a obtenu de meilleurs résultats mais a nécessité un travail plus conséquent avec la création d'un nouveau modèle. Il est donc difficile de déterminer quelle RTO, de la SNOMED CT et du NCIt, est la plus adaptée pour utiliser conjointement la CIM-10 et la CIM-O3.

Les deux études présentées dans cette section ont indirectement mis en évidence des défauts de qualité au sein de la RTO utilisée comme base du modèle d'intégration ou comme support. Plus précisément, les facteurs d'exhaustivité de la couverture et d'exactitude définis par Zhu *et al.* [114] sont concernés. En effet, nous avons observé l'absence de certains concepts dans le NCIt lorsque des concepts CIM-10 n'ont pas pu être intégrés au modèle car les diagnostics correspondants n'existaient pas dans le NCIt. Nous avons également identifié qu'un nombre important de concepts SNOMED CT étaient décrits comme enfants d'un même concept alors qu'ils ne devraient pas l'être. Par ailleurs, via les codes CIM-10 instanciant plusieurs concepts de diagnostics, nous avons noté que les concepts NCIt correspondants étaient reliés à de multiples sites anatomiques. Par exemple, le concept CIM-10 Malignant neoplasm : Caecum (C18.0) a été mappé au concept NCIt Malignant Cecum Neoplasm (C9329) qui est relié aux cinq sites anatomiques suivants : Gastrointestinal System, Cecum, Colon, Intestine et Colorectal Region. Le concept CIM-10 C18.0 instanciait finalement les concepts Malignant, primary site - Cecum Reflexive part, Malignant, primary site - Colon Reflexive part et Malignant, primary site - Colorectal Region Reflexive part dans le modèle créé. Ce type de situation révèle deux sortes de problèmes au sein du NCIt :

— l'absence de certaines relations. En effet, Cecum et Colon devraient être associés à Colorectal Region via une relation de type *part_ of*.

— des incohérences structurelles dues à l'héritage multiple qui est parfois utilisé pour refléter la polysémie des concepts. Ce problème typique des RTOs biomédicales a été qualifié d'*is_ a overloading* par Guarino [188] et illustré dans le NCIt par Kumar et Smith [189]. Ainsi, Malignant Cecum Neoplasm est décrit dans le NCIt comme sous-concept de Cecum Neoplasm et de Malignant Colon Neoplasm pour préciser que c'est un type de néoplasme du cæcum et un type de cancer du colon. C'est pour cette raison qu'il a pour site anatomique Colon, ce qui ne devrait pas être le cas car le cæcum n'est ni une partie, ni un sous-concept du colon.

3.3 Synthèse et perspectives

Trois **constats** ressortent de ce chapitre : (i) des techniques de différents types peuvent être utilisées pour établir des correspondances entre RTOs, (ii) les ressources externes sont utiles à l'alignement et à l'intégration mais elles peuvent nécessiter de réaliser des étapes supplémentaires, et (iii) l'utilisation conjointe de RTOs permet de faire indirectement de l'évaluation.

Dans les quatre travaux présentés dans ce chapitre, nous avons utilisé plusieurs techniques afin d'établir des correspondances entre RTOs. Pour l'alignement, nous avons exploité une technique *informal resource-based* pour filtrer les mappings obtenus par notre méthode *language-based*. Dans le cas de l'intégration, c'est une technique *informal resource-based* qui a permis d'obtenir les mappings avec une RTO intermédiaire tandis que la technique *formal resource-based* a servi à mettre en correspondance les RTOs à intégrer. Nous avons utilisé ces différentes techniques de manière séquentielle. Certains travaux participant aux campagnes OAEI offrent la possibilité d'exécuter en parallèle différentes techniques puis de fusionner leurs résultats [157]. Par exemple, SAMBO propose des techniques de types *string-based*, *language-based*, *taxonomy-based*, *informal resource-based* et *instance-based* qui peuvent être combinées dans l'ordre choisi par l'utilisateur du système [190]. Il serait intéressant d'explorer cette possibilité.

Les ressources externes exploitées dans nos études ont été utiles dans de multiples situations : pour filtrer des mappings, pour évaluer des mappings, pour établir des mappings entre les RTOs à intégrer et pour servir soit de base au modèle d'intégration, soit de support. Cependant, lors du processus d'intégration, il a été nécessaire de mettre en œuvre des étapes supplémentaires afin de corriger certaines connaissances décrites dans ces ressources. Ainsi, nous avons dû modifier l'organisation des concepts de maladie et de morphologie de la RTO utilisée comme base du modèle d'intégration et compléter la description des concepts anatomiques. Dans la seconde étude, nous avons filtré les mappings fournis par une ressource informelle afin d'éliminer ceux qui associaient des concepts appartenant à des hiérarchies différentes ou dont le comportement tumoral n'était pas le même.

Finalement, nos travaux ont montré que lors des processus d'alignement et d'intégration, il était possible d'évaluer indirectement la qualité des ressources externes exploitées, et potentiellement de l'améliorer [191]. En effet, nous avons identifié des problèmes relevant des critères d'inexactitude et d'exhaustivité de la couverture définis par Zhu *et al.* [114]. Plus précisément, nous avons constaté l'absence de relations associatives et de concepts mais aussi la présence de relations hiérarchiques inappropriées. Il est intéressant de noter que certaines de ces erreurs ont pu être identifiées automatiquement, par exemple lorsque des concepts des RTOs à intégrer ne pouvaient être associées à aucun concept de la RTO utilisée comme base au modèle d'intégration.

En termes de **perspectives**, nous envisageons de participer aux tâches dédiées aux RTOs biomédicales lors des prochains défis de l'OAEI. Pour l'instant, nous avons développé des méthodes pour aligner ou intégrer des RTOs biomédicales particulières alors qu'il faudrait proposer des méthodes génériques. Jean Noël essaie actuellement d'implémenter le cadre conceptuel défini dans [96] pour pouvoir intégrer n'importe quelles RTO biomédicales décrivant des notions différentes mais pouvant être mises en relation via une RTO de support. C'est une opportunité idéale pour poursuivre nos recherches sur la découverte de mappings complexes qui a été peu étudiée jusqu'ici. Un autre champ à approfondir concerne le choix de la RTO de support. Dans nos travaux, nous l'avons fait manuellement mais des chercheurs ont tenté d'automatiser

ce processus dans le cadre de la campagne OAEI 2011 [192] et dans celle de 2013 avec un travail plus spécifiquement appliqué au domaine biomédical [193]. Notons qu'il serait également utile de pouvoir découvrir automatiquement les relations permettant de mettre en correspondance les différentes types de concepts décrits dans les RTOs à intégrer (*e.g.*, *associated_ morphology* entre un concept CIM-10 de maladie et un concept CIM-O3 de morphologie). Enfin, la figure 3.1 montre qu'il reste des techniques que nous n'avons pas encore explorées alors que les RTOs biomédicales disposent des caractéristiques nécessaires à leur mise en œuvre. En particulier, les techniques de types *string-based* et *taxonomy-based* sont tout à fait adaptées aux RTOs biomédicales, vu leur composante terminologique et leur structuration hiérarchique.

CHAPITRE 4

CONCEPTION DE RESSOURCES TERMINO-ONTOLOGIQUES

Malgré la multiplicité des RTOs qui existent dans le domaine biomédical, des besoins spécifiques peuvent nécessiter de créer une nouvelle RTO pour décrire les connaissances lié à un thème donné de manière plus détaillée. Comme je l'ai mentionné dans la section 1.4, de nombreuses méthodologies de conception d'ontologies ont été proposées autour des années 2000 (*e.g.*, [59, 60, 61] et une revue de la littérature de celles-ci est présentée dans [58]). La conception d'ontologies pouvant s'avérer particulièrement fastidieuse, notamment dans des domaines vastes et complexes tels que la santé, des méthodes ont été mises en œuvre pour automatiser certaines étapes du développement d'ontologies (*ontology learning*). En particulier, l'étape d'acquisition de connaissances (définie comme activité de support dans la figure 1.3) a fait l'objet d'un grand nombre de travaux de recherche. Dans ce cadre, les sources de connaissances exploitables peuvent être disponibles sous différentes formes [194, 195, 196] : non structurée (*i.e.*, en texte libre), semi-structurée (*e.g.*, décrites en XML, présentes sur le Web) ou structurée (*e.g.*, RTOs, données stockées dans des bases de données relationnelles). Dans sa revue de la littérature, Zhou a classé les recherches effectuées sur cette thématique suivant plusieurs axes [195], incluant les différents **types de connaissances** pouvant être extraits, les **stratégies suivies** et les **techniques mises en œuvre**.

Buitelaar *et al.* ont défini les types de connaissances, par ordre croissant de difficulté à acquérir, comme suit [197] :

— les termes,

— les synonymes, à savoir les variants sémantiques des termes, ceux-ci pouvant être des variants linguistiques (*i.e.*, des traductions des termes),

— les concepts, que les auteurs décrivent selon : (i) leur intension, caractérisée par leur définition logique ou leur définition textuelle, (ii) leur extension, définie par leurs instances, et (iii) les termes se référant aux concepts,

— les relations hiérarchiques entre concepts,

— les relations associatives entre concepts,

— les axiomes.

Les stratégies définies par Zhou sont les approches ascendantes (*bottom-up*), descendantes (*top-down*) ou hybrides. L'auteure indique que le domaine pour lequel une RTO doit être créée peut orienter le choix de la stratégie à suivre. Par exemple, les domaines émergents pour lesquels aucune RTO n'a été développée devraient tirer le meilleur profit d'une stratégie ascendante pour extraire les connaissances d'intérêt, à partir de textes notamment. À l'inverse, un domaine interdisciplinaire qui est de fait sujet à des ambiguïtés liées à des usages différents des termes dans chaque discipline gagne *a priori* à adopter une stratégie descendante en définissant tout d'abord les connaissances de haut niveau qui pourront ensuite être raffinées.

Enfin, Zhou a classé les techniques proposées pour acquérir des connaissances dans les trois catégories suivantes : les techniques statistiques, les techniques basées sur des règles (également qualifiées de linguistiques dans d'autres revues [197, 198]) et les techniques hybrides qui mixent les deux autres types de techniques. Parmi les travaux recensés dans ces différentes revues, on peut citer les techniques basées sur des ressources (ou lexiques) sémantiques externes, sur de la fouille de texte, sur le regroupement de données (*clustering*), sur différentes sortes de méthodes d'apprentissage ou encore sur des patrons lexico-syntaxiques. Il est d'ailleurs fréquent que les travaux présentés combinent plusieurs de ces techniques, certaines étant plus adaptées à un type donné de connaissances. Dans la revue plus récente de Wong *et al.*, les auteurs ont proposé une classification des techniques existantes au regard des tâches permettant d'acquérir les différents types de connaissances introduits ci-dessus [198]. Ils ont par ailleurs identifié une catégorie supplémentaire de techniques ; celles basées sur la logique, et plus précisément sur la déduction ou l'induction.

Dans ce chapitre, je présente deux travaux ayant nécessité le développement d'une RTO. Le contexte de ces deux études étant différent, cela illustre les raisons qui motivent à s'orienter vers une stratégie ou une autre. D'un côté, le domaine d'application est une maladie donnée pour laquelle aucune RTO n'a été conçue spécifiquement alors qu'un niveau de détails très fin est requis en terme de représentation. Le but étant de réaliser un moteur de recherche pour accéder à des documents textuels dans plusieurs langues, la stratégie adoptée dans ce travail est ascendante avec une acquisition des connaissances à partir de ces textes et leur structuration grâce à une ressource externe décrivant le domaine biomédical plus globalement. La seconde étude s'intéresse à un domaine interdisciplinaire pour lequel la plupart des connaissances à représenter sont disponibles de manière parcellaire dans différentes RTOs. Dans ce cas, la problématique est d'intégrer les RTOs jugées pertinentes et d'acquérir les connaissances manquantes pour couvrir le domaine d'intérêt dans son ensemble. La stratégie suivie est donc hybride puisqu'elle repose sur l'identification des concepts de haut niveau dans les RTOs appropriées, l'intégration de ces RTOs et l'exploitation de textes pour extraire les connaissances spécifiques au domaine.

Dans la section 4.1, je décris une partie du travail de thèse de Khadim Dramé qui visait à concevoir une RTO bilingue à partir de documents textuels et d'une ressource externe afin d'automatiser le processus de conception (a,b). J'introduis ensuite les réalisations de Georgeta Bordea, une post-doctorante que j'encadre dans le cadre du projet ANR MIAM. Elle a tout d'abord constitué un corpus d'articles décrivant des interactions médicament-aliment puisqu'une recherche classique dans la base de données bibliographiques MEDLINE ne permettait pas de récupérer l'ensemble des articles pertinents sur ce sujet (c,d). Dans la section 4.2, je présente la RTO que nous modélisons actuellement pour représenter les interactions médicament-aliment en suivant un processus d'intégration (e).

(a) Khadim Dramé, Gayo Diallo, Fleur Mougin. Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus. Proceedings of the 24th Conference of Medical Informatics in Europe, pages 179-183, 2012

(b) Khadim Dramé, Gayo Diallo, Fleur Delva, Jean-François Dartigues, Evelyne Mouillet, Roger Salamon, Fleur Mougin. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: an application to Alzheimer's disease. Journal of Biomedical Informatics, 48:171 – 182, 2014

(c) Georgeta Bordea, Frantz Thiessard, Thierry Hamon, Fleur Mougin. Automatic query selection for acquisition and discovery of food-drug interactions. Proceedings of the 9th Conference and Labs of the Evaluation Forum, pages 115-120, 2018

(d) *Georgeta Bordea, Tsanta Randriatsitohaina, Natalia Grabar, Fleur Mougin, Thierry Hamon. Query selection methods for automated corpora construction with a use case in food-drug interactions. Soumis au ACL BioNLP workshop, 2019*

(e) *Georgeta Bordea, Romain Griffier, Vianney Jouhet, Fleur Mougin. FIDEO: an ontology to represent food-drug interactions. Article en cours de rédaction, 2019*

4.1 Conception d'une ressource termino-ontologique à partir de textes et d'une ressource externe

4.1.1 Contexte médical : la maladie d'Alzheimer

Parmi les nombreux défis à relever en santé publique, la maladie d'Alzheimer et les syndromes apparentés constituent un enjeu majeur. En 2008, le nombre de patients atteints par l'une de ces maladies était de 850 000 en France et ce nombre devrait dépasser 2 millions en 2040 [1]. Le plan Alzheimer 2008-2012 incluait 44 mesures, dont la 32ème mesure avait pour but de former les professionnels de santé à l'épidémiologie clinique afin qu'ils s'impliquent dans la recherche et incitent leurs patients à participer à des études cliniques. Aujourd'hui, c'est plus largement les maladies neurodégénératives qui font l'objet d'un plan, le plan MND 2014-2019 [2], au sein duquel la maladie d'Alzheimer a une place prépondérante.

Le bulletin bibliographique mensuel BiblioDémences [3] s'inscrit dans les objectifs de la mesure 32 du plan Alzheimer 2008-2012 et dans la grande priorité "Développer et coordonner la recherche" du plan MND 2014-2019. Lancé en 2004 par l'équipe Épidémiologie et Neuropsychologie du Vieillissement Cérébral du centre de recherche Inserm Bordeaux Population Health, il fournit une analyse critique en français d'articles scientifiques sur la maladie d'Alzheimer et les syndromes apparentés. L'objectif est de rendre disponible aux professionnels de santé et aux décideurs publics l'avis d'un expert du domaine sur la littérature internationale incontournable concernant ces maladies. En pratique, une documentaliste récupère chaque mois les articles scientifiques potentiellement d'intérêt (notons que les articles trop théoriques sont ignorés car BiblioDémences a vocation à aider les professionnels de santé et le personnel soignant dans leur pratique au quotidien) parmi lesquels le comité éditorial de ce bulletin en sélectionne 30 à 50. Une analyse critique de chacun de ces articles est alors réalisée par un expert du domaine puis l'article et l'analyse sont enregistrés dans une base de données, nommée BiblioDem. Les analyses critiques des 10 à 15 articles jugés les plus pertinents sont publiées dans le bulletin BiblioDémences.

Le moteur de recherche qui permettait d'interroger BiblioDem offrait des fonctionnalités limitées puisque seuls les articles dont le titre ou le résumé contenant une occurrence exacte du terme recherché étaient retournés. De plus, la variété des utilisateurs n'était pas prise en compte alors que des termes scientifiques et des termes utilisés par le grand public pouvaient être employés pour faire une recherche d'information. C'est dans ce cadre qu'avec mon collègue Gayo Diallo, nous avons obtenu un financement de 100 000€ par la fondation Plan Alzheimer pour le projet SemBiP (Semantic BiblioDem Portal) dont le but était de créer un portail sémantique basé sur une RTO fournissant un accès rapide et efficace aux articles de la base de données BiblioDem. Nous avons ainsi recruté un doctorant, Khadim Dramé, qui a effectué sa thèse entre 2011 et 2014 sur ce sujet. Les travaux qu'il a réalisés pour la conception de la RTO, nommée OntoAD, sont présentés dans cette section [84, 85, 86].

1. http://archives.gouvernement.fr/fillon_version2/gouvernement/le-plan-alzheimer-2008-2012.html
2. https://www.gouvernement.fr/action/le-plan-maladies-neuro-degeneratives-2014-2019
3. http://sites.isped.u-bordeaux2.fr/bibliodem/index.aspx

4.1.2 Étapes de la conception d'OntoAD

Je présente ici le processus mis en œuvre pour développer OntoAD suivant la méthodologie de conception METHONTOLOGY [59] et ses sept étapes (introduites dans la section 1.4). La figure 4.1 illustre ce processus.

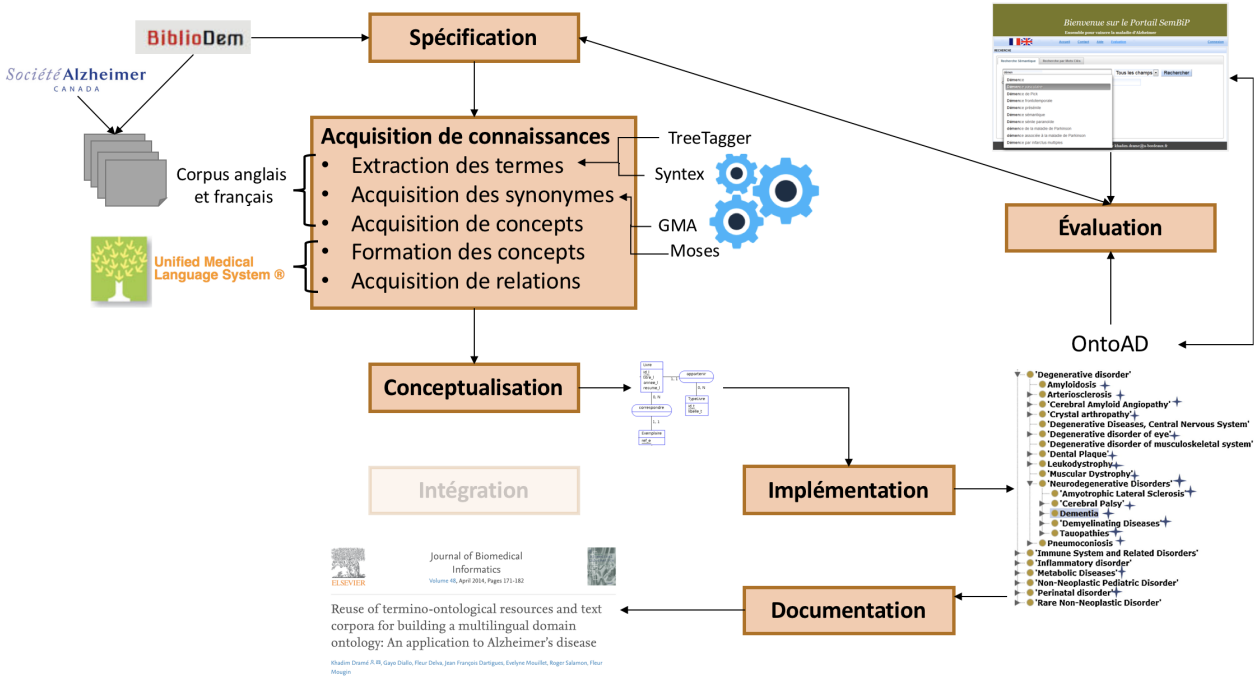

FIGURE 4.1 – Processus de conception d'OntoAD. Les étapes sont représentées dans des rectangles orange (l'étape d'intégration apparaît en transparence car elle n'a pas été mise en œuvre dans ce travail).

Spécification

Les objectifs du projet SemBiP étaient les suivants : (i) fournir une RTO multilingue (en français et en anglais, pour commencer) sur la maladie d'Alzheimer et les syndromes apparentés pour indexer sémantiquement les articles de BiblioDem, mais aussi à des fins pédagogiques, (ii) implémenter un portail sémantique permettant d'effectuer une recherche dans plusieurs langues sur des documents de référence au sujet de ces maladies, (iii) prendre en compte la spécificité et la variété des utilisateurs du portail en proposant des fonctionnalités de recherche adaptées au profil de l'utilisateur. Ainsi, OntoAD a été conçue pour recenser les connaissances sur la maladie d'Alzheimer et les syndromes apparentés et pour servir de socle à un moteur de recherche d'information sémantique. Dans ce cadre, une RTO décrivant les concepts et les relations entre eux de manière formelle (*light-weight ontology*) était suffisante [199].

Les utilisateurs d'OntoAD étaient de deux types : 1) les personnes en charge de l'indexation des documents de BiblioDem et de l'implémentation du portail sémantique (*i.e.*, Khadim), 2) les utilisateurs finaux de SemBiP, à savoir les chercheurs alimentant BiblioDem, le personnel soignant, les étudiants en médecine et les décideurs publics.

Acquisition de connaissances

Pour acquérir les connaissances à représenter dans OntoAD, deux sources de connaissances ont été exploitées : des corpus de textes concernant la maladie d'Alzheimer et les syndromes apparentés pour identifier les termes pertinents du domaine et l'UMLS pour former les concepts à partir de ces termes et décrire des relations entre les concepts. Par ailleurs, des outils de traitement automatique des langues ont été utilisés lors des deux étapes suivantes : l'acquisition des termes du domaine et l'extraction des synonymes au sein des textes.

Extraction des termes. Khadim a tout d'abord constitué deux corpus à partir de BiblioDem. En juillet 2013, cette base de données contenait 1556 articles pour lesquels sont enregistrés le titre, les auteurs, le résumé en anglais (*i.e.*, l'*abstract*), l'analyse critique en français et le nom de l'expert qui l'a réalisée. Chaque analyse critique est composée d'une traduction en français du titre de l'article, d'un résumé (qui n'est pas une traduction de l'abstract) et des commentaires de l'expert. Ainsi, nous avons constitué un corpus anglais contenant l'abstract des 1556 articles et un corpus français comprenant l'analyse critique de ces articles.

Pour l'extraction des termes du domaine, l'analyseur syntaxique Syntex [200] a été utilisé sur ces deux corpus. Cet outil de traitement automatique des langues prend en entrée un corpus de texte étiqueté (étiquetage réalisé avec TreeTagger [4] pour ce travail) et fournit en sortie un réseau terminologique des syntagmes nominaux et verbaux trouvés dans les textes et pour lesquels des statistiques, telles que la fréquence des termes, sont également fournies. Khadim a alors appliqué un filtre sur les termes candidats obtenus avec Syntex pour éliminer les mots vides [5], les termes constitués uniquement de chiffres et ceux qui apparaissaient moins de sept fois. Ce seuil a été choisi par les experts du domaine et, bien que le filtre ait impliqué la suppression de termes potentiellement très spécifiques du domaine, il permettait de récupérer les termes importants sans pour autant résulter en un trop grand nombre de termes.

Regroupement des termes en concepts et acquisition de relations. Cette tâche visait à constituer une première ébauche de la RTO à partir des termes sélectionnés. Ainsi, grâce à l'UMLS, les termes synonymes ont été regroupés au sein de concepts et les relations de subsomption et associatives (*e.g.*, *may_treat*, *cause_of*, *anatomical_part_of*) permettant de relier ces concepts ont été collectées. Lorsque des concepts étaient reliés de manière indirecte, les concepts intermédiaires dans l'UMLS ont été récupérés afin d'obtenir une structuration plus complète. Par ailleurs, l'ensemble des concepts descendants de neuf concepts de haut niveau jugés pertinents par les experts pour décrire le domaine ont été extraits (*e.g.*, Impaired cognition (C0338656), Frontotemporal dementia (C0338451), Mini-mental state examination (C0451306)). Enfin, les types sémantiques de l'UMLS ont été intégrés dans OntoAD pour organiser tous ces concepts (y compris ceux qui n'étaient reliés à aucun autre concept).

Malgré sa richesse, le metathesaurus de l'UMLS présente des situations problématiques, comme illustré dans la section 2.1. En particulier, lors de ce travail, nous avons identifié la présence de relations erronées (générant des cycles) et de relations redondantes. Pour pallier ces problèmes, nous avons défini les règles suivantes inspirées du travail de Christment *et al.* [201] :

4. `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/`
5. `https://nlp.cs.nyu.edu/GMA/docs/resources.html`

— si une relation de subsomption existait entre deux concepts, toute autre relation entre ces deux concepts a été supprimée,

— si deux concepts étaient reliés par plusieurs relations, elles-mêmes reliées hiérarchiquement, seule la relation la plus spécifique a été conservée,

— les relations pouvant être déduites par transitivité ont été éliminées pour ne pas alourdir OntoAD,

— si un cycle impliquant des concepts reliés hiérarchiquement existait, la relation étant à l'origine de ce cycle a été supprimée,

— si le terme préféré d'un concept donné C_1 apparaissait parmi les termes d'un autre concept C_2, il a été retiré des termes de C_2.

Ces règles ont été appliquées automatiquement, hormis celle concernant les cycles qui a nécessité une intervention manuelle pour déterminer la relation hiérarchique erronée.

Acquisition des synonymes. Pour développer l'aspect multilingue d'OntoAD, Khadim a constitué des corpus parallèles. Le premier contenait le titre des articles en anglais et en français, exploitant ainsi les traductions du titre faites par les experts. Pour compléter ce corpus et acquérir un vocabulaire moins spécialisé, des textes issus du site Internet bilingue anglais-français de la société Alzheimer Canada [6] ont été collectés. Sur la partie grand public, la plupart des pages Web sont disponibles en anglais et en français et sont généralement des traductions l'une de l'autre. Ainsi, 705 paires de documents parallèles ont pu être récupérées en plus du corpus parallèle des titres.

Grâce à ces corpus, les concepts d'OntoAD ont été enrichis avec des termes français car l'UMLS contient principalement des termes anglais. Pour cela, un lexique bilingue contenant les titres des articles en anglais et leur traduction en français a été créé et fusionné avec un lexique bilingue obtenu à partir de l'UMLS. Les corpus parallèles ont alors été traités avec l'outil GMA (Geometrical Mapping and Alignment) [7], qui permet d'apparier des phrases en anglais et en français en exploitant les cognats, un lexique bilingue et la similarité entre mots basée sur la plus longue séquence de caractères commune [202]. À partir des 10 000 paires de phrases obtenues, les termes ont été alignés en prenant en compte le nombre d'occurrences entre deux termes de langues différentes et leur similarité morphosyntaxique [84]. Bien qu'efficace sur le corpus parallèle de titres (précision de 73%), notre méthode s'est avérée moins performante sur les corpus constitués à partir du site Internet de la société Alzheimer Canada. Khadim a donc appliqué dans un deuxième temps le logiciel Moses, qui effectue de la traduction automatique statistique de termes composés de plusieurs mots et fournit une probabilité de traduction associée [203].

Acquisition de concepts. Finalement, des concepts qui n'existaient pas dans l'UMLS ont été ajoutés à OntoAD. Pour cela, les termes anglais non trouvés dans l'UMLS ont été considérés comme candidats (car nous avons décidé que tout concept d'OntoAD devait avoir au moins un terme en anglais). Ceux qui ont été retenus par les experts ont été regroupés en concepts quand ils étaient synonymes (anglais ou français) et reliés aux concepts issus de l'UMLS via des relations de dépendance (*i.e.*, tête et expansion) générées par Syntex, s'il en existait. Par

6. `http://alzheimer.ca/fr/Home`
7. `http://nlp.cs.nyu.edu/GMA/`

exemple, le nouveau concept Severe Alzheimer disease (AD000390) a été intégré dans OntoAD comme sous-concept de Alzheimer disease (C0002395).

Conceptualisation

La figure 4.2 présente le modèle conceptuel suivant lequel les concepts, relations et termes ont été représentés dans OntoAD. Chaque concept est ainsi identifié par un URI (Uniform Resource Identifier), un terme préféré en anglais et éventuellement un terme préféré en français, zéro à plusieurs synonymes en anglais et/ou en français et zéro à plusieurs définitions textuelles. Pour chaque terme et chaque définition, la langue dans laquelle ils sont définis est précisée. Enfin, les concepts sont reliés entre eux via des relations hiérarchiques et/ou associatives, le nom de la relation devant être spécifié.

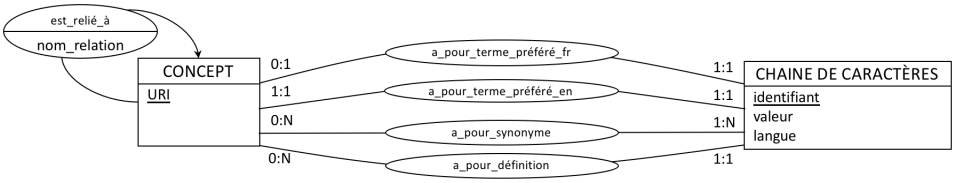

FIGURE 4.2 – Modèle conceptuel d'OntoAD.

Intégration

Contrairement à ce qui est préconisé dans METHONTOLOGY, OntoAD n'a pas été intégrée à une ontologie de haut niveau car son utilisation dans le cadre de SemBiP ne rendait pas cette étape indispensable.

Implémentation

Khadim a décrit OntoAD en combinant OWL et SKOS, ce dernier étant particulièrement bien adapté pour supporter le multilinguisme [204]. Plus précisément, les concepts et les relations ainsi que leur(s) caractéristique(s) (*e.g.*, symétrique, transitive, inverse) ont été représentés en OWL tandis que les termes (`skos:prefLabel` pour les termes préférés et `skos:altLabel` pour les synonymes) et les définitions (`skos:definition`) ont été décrits en SKOS.

Évaluation

Cette étape englobe deux aspects : la validation et l'évaluation [27]. La validation a été réalisée par deux spécialistes du domaine de l'équipe alimentant BiblioDem sur les éléments suivants d'OntoAD :

— les concepts trouvés dans l'UMLS avec leurs termes associés et leur contexte d'apparition dans le corpus constitué à partir de BiblioDem,

— une liste des types de relations associatives extraites de l'UMLS,

— les termes anglais non trouvés dans l'UMLS et leur traduction en français, si elle existait.

L'adéquation d'OntoAD aux besoins exprimés lors de la spécification a été évaluée partiellement au travers du portail sémantique SemBiP. Tout d'abord, Khadim a mis en œuvre une méthode d'indexation sémantique qui comprenait deux étapes : 1) une phase de repérage des concepts d'OntoAD dans des corpus de documents en privilégiant les plus spécifiques, 2) une méthode de désambiguïsation de concepts basée sur la similarité sémantique. Il a ensuite créé un prototype de SemBiP, qui permet d'accéder facilement à une analyse critique de la littérature mondiale de référence sur la maladie d'Alzheimer et les syndromes apparentés. Pour la phase de recherche d'information, Khadim a appliqué une technique d'expansion de requêtes basée sur la hiérarchie des concepts d'OntoAD. Plus précisément, la requête était complétée avec les concepts enfants sémantiquement proches des termes la composant [87]. Pour évaluer l'intérêt des fonctionnalités mises en œuvre dans SemBiP, Khadim a établi un questionnaire avec huit questions (dont sept nécessitaient d'attribuer un score de 1 à 5 et la dernière était une question ouverte) portant sur le moteur de recherche, et indirectement sur OntoAD. Bien que cette évaluation soit limitée car seulement huit utilisateurs ont répondu, leur satisfaction globale était positive.

Documentation

Fernández-López *et al.* ont précisé que chaque étape du développement doit être documentée [59]. Nous n'avons pas réalisé un document explicatif de chaque étape mais le descriptif du projet SemBiP que nous avons soumis pour obtenir un financement détaille bien les aspects liés à l'étape de spécification. Par ailleurs, la publication présentant OntoAD et la manière dont elle a été conçue documente chaque point décrit dans les différentes étapes [86], tout comme le manuscrit de thèse de Khadim [83].

4.1.3 OntoAD en chiffres

Lors de l'étape d'extraction de termes, 2 916 termes anglais et 3 152 termes français ont été retenus. La recherche de ces termes dans l'UMLS a résulté en 3 871 concepts distincts dont 2 421 ont été validés par les experts. Pour les structurer, 2 905 concepts ont été ajoutés. En plus des relations hiérarchiques de l'UMLS, 279 relations obtenues grâce à Syntex ont été intégrées. À partir des 1527 paires de termes anglais-français validées par les experts, 608 termes français ont été inclus comme synonymes de concepts. Enfin, 439 concepts non présents dans l'UMLS ont été ajoutés à OntoAD.

OntoAD est disponible sur BioPortal [8], un portail Web qui fournit un accès à plus de 750 RTOs biomédicales [205], et contient au total :

— 5 765 concepts, dont 3 283 (56%) ont un synonyme en français,

— 7 499 relations hiérarchiques entre concepts,

— 178 types de relations associatives,

— 10 889 relations associatives,

— 35 855 termes (23 934 termes anglais et 11 921 termes français),

— aucun axiome.

8. `http://bioportal.bioontology.org/ontologies/ONTOAD`

4.1.4 Conclusions

Au regard de la classification de Wong *et al.* [198], les techniques utilisées pour concevoir OntoAD sont essentiellement linguistiques. Plus précisément, une étape préliminaire d'étiquetage morphosyntaxique effectuée avec TreeTagger prépare le corpus pour que Syntex puisse extraire les termes du domaine grâce à une technique basée sur l'analyse syntaxique. Celle-ci permet également de construire une hiérarchie entre ces termes grâce aux relations de dépendance. Ensuite, l'utilisation d'un lexique sémantique (*i.e.*, l'UMLS) permet de former les concepts à partir des termes obtenus avec Syntex, de construire une hiérarchie entre ces concepts et d'obtenir des relations associatives entre eux. La traduction sert à acquérir des synonymes (correspondant à des variants linguistiques), type de connaissances qui apparaît dans la classification de Buitelaar *et al.* [197], contrairement à celle de Wong *et al.* Cette étape repose d'une part sur des techniques linguistiques avec un algorithme de reconnaissance de formes, l'utilisation d'un lexique bilingue et d'un lexique sémantique, et d'autre part sur une technique de traduction automatique statistique.

OntoAD est une RTO que l'on peut qualifier d'*ontologie légère* puisqu'elle contient des concepts organisés suivant des relations de subsomption et associatives, mais pas d'axiomes. Cependant, vu qu'elle a été utilisée pour l'indexation de documents et la recherche d'information sémantique, ce niveau de formalisation était suffisant. Comme souligné par Aussenac-Gilles *et al.*, la forte composante terminologique d'OntoAD, notamment son caractère bilingue, constitue un réel avantage pour ce type d'utilisation [206]. La méthodologie Terminae définie par ces auteures a pour but de concevoir une ontologie légère par l'acquisition de connaissances à partir de textes. L'approche que nous avons suivie pour construire OntoAD est conforme aux principes de Terminae puisqu'elle a été créée à partir de textes grâce à des outils de traitement automatique des langues (notamment TreeTagger et Syntex, qui sont cités dans [206]) mais aussi en réutilisant une ressource externe. De plus, le développement de cette RTO a été fait de manière supervisée puisque les experts sont intervenus pour valider son contenu. En revanche, cette RTO n'est pas rattachée à une ontologie de haut niveau alors que Terminae et METHONTOLOGY le recommandent. C'est également une préconisation de la communauté développant des ontologies formelles, en particulier le consortium OBO Foundry qui héberge des RTOs biomédicales utilisant BFO (Basic Formal Ontology) [207] comme ontologie de haut niveau [208]. Ainsi, une perspective intéressante serait d'intégrer OntoAD avec les ontologies développées par Hastings *et al.* [209] car elles décrivent le domaine plus vaste des maladies mentales et sont positionnées par rapport à BFO.

4.2 Conception d'une ressource termino-ontologique par intégration

4.2.1 Contexte médical : les interactions entre médicaments et aliments

Le projet ANR PRCE MIAM (Maladies, Interactions Aliments-Médicaments / 2017-2020) vise à identifier et modéliser les interactions entre aliments et médicaments. Il est coordonné par Thierry Hamon et est constitué des équipes de recherche ILES (Information Langue Écrite et Signée) du LIMSI, ERIAS et MABioVis (Modèles et Algorithmes pour la Bioinformatique et la Visualisation d'informations) de Bordeaux, STL (Savoirs, Textes, Langage) de Lille, le centre de pharmacovigilance de Bordeaux, l'entreprise Antidot [9] et le CNHIM (Centre National Hospitalier d'Information sur le Médicament). Ce projet s'intéresse aux problèmes qui peuvent émerger lorsqu'une interaction survient entre un aliment et un médicament. En pratique, la consommation d'un aliment associée à la prise d'un médicament peut avoir trois effets : (i) l'atténuation ou la suppression de l'effet d'un médicament, (ii) l'augmentation ou la diminution de l'effet d'un médicament, et (iii) l'apparition d'effets indésirables encore inconnus. Ces informations, quand elles sont disponibles, sont décrites dans des publications scientifiques ou dans des bases de connaissances de manière non structurée. DrugBank notamment, qui est pourtant mise à disposition sous la forme de données liées, décrit des interactions aliment-médicament mais cette information est fournie en texte libre et donc difficilement exploitable telle quelle de façon automatique. Par exemple, les interactions de la substance chimique lévothyroxine (contenue notamment dans le médicament Levothyrox®) avec des aliments sont décrites comme suit : "*Absorption increased in fasting state and decreased in malabsorption states. Consistent administration in relation to meals is recommended. No iron or calium carbonate within 4 hours of taking this medication. Oral administration with infant soybean formula, soybean flour, cotton seed meal, walnuts, foods containing large amounts of fiber, ferrous sulfate, and antacids may decrease drug absorption. Take 30-60 minutes before breakfast*" [10]. La complexité posée par ces interactions est bien illustrée au travers de cet exemple, en particulier :

— la présence de fautes d'orthographe (ici, "calium carbonate" devrait en fait être "calcium carbonate"),

— le fait que les interactions peuvent être décrites entre un aliment et un médicament, mais aussi entre un aliment et une substance chimique ou encore entre un médicament et une classe d'aliments (ceux contenant une grande quantité de fibres dans notre exemple),

— le mécanisme mis en jeu peut être différent en fonction de l'état de la personne qui consomme le médicament (dans l'exemple de la lévothyroxine, son absorption est différente si la personne est à jeun ou sujette à des troubles digestifs),

— une composante temporelle à prendre en compte pour décrire le délai entre la prise du médicament et l'alimentation, mais aussi le moment de la journée où il doit être pris.

Dans ce cadre, les objectifs du projet MIAM sont de :

— collecter les données non structurées sur les interactions entre un aliment et un médicament dans les articles scientifiques, les bases de connaissances et les dépôts de données liées,

9. `http://www.antidot.net`
10. `https://www.drugbank.ca/drugs/DB00451`

— extraire les entités d'intérêt (*e.g.*, médicament, aliment, maladie) au sein des textes ainsi que les relations existant entre elles,

— identifier la fiabilité de l'information extraite,

— représenter les connaissances liées aux entités d'intérêt et en particulier les interactions entre elles,

— alimenter Thériaque®, qui est une base de données sur les médicaments disponibles en France développée par le CNHIM [11], avec les résultats obtenus dans le cadre du projet.

Notre équipe (ERIAS) a en charge de représenter les connaissances liées aux interactions entre aliments et médicaments. Georgeta Bordea nous a rejoint en juin 2017 pour effectuer son post-doctorat sur cette problématique. Dans cette section, je détaille la méthodologie que nous avons suivie pour concevoir FIDEO (Food Interacting with Drug Evidence Ontology) qui est toujours en cours de développement.

4.2.2 Étapes de la conception de FIDEO

Comme pour OntoAD, je présente le processus de conception de FIDEO d'après les sept étapes de METHONTOLOGY [59].

Spécification

Les besoins pour lesquels FIDEO a dû être construite ont été exprimés initialement lors de la rédaction du projet. Elle doit permettre de représenter les connaissances nécessaires à la description des interactions entre un aliment et un médicament. Grâce à ce modèle, l'intégration des informations en lien avec ces interactions dans Thériaque sera facilitée et structurée afin de favoriser leur réutilisation. De plus, quand les interactions impliquent une catégorie d'aliments (ou une classe de médicaments), il est nécessaire de pouvoir inférer que des interactions potentielles peuvent se produire lors de la consommation de chacun des aliments de la catégorie concernée (ou lors de la prise de chaque médicament de la classe concernée). Enfin, FIDEO doit pouvoir être utilisée pour identifier des interactions potentielles entre un aliment et un médicament dans les futurs articles scientifiques pour continuer à alimenter Thériaque au fur et à mesure que de nouvelles connaissances apparaîtront. Étant donné les utilisations envisagées de FIDEO et en particulier le besoin d'inférences, son niveau de formalisation doit être élevé.

Les utilisateurs de FIDEO sont de différents profils : 1) les chercheurs en traitement automatique des langues qui vont l'utiliser comme ressource sémantique dans leurs méthodes d'extraction d'information, 2) les personnes du CNHIM qui administrent la base de données Thériaque et le moteur de recherche associé (qui pourrait reposer sur FIDEO), 3) les utilisateurs finaux de Thériaque via le moteur de recherche, à savoir les professionnels de santé et les patients.

Acquisition de connaissances

Corpus. Pour recenser les connaissances à décrire dans FIDEO, Georgeta a tout d'abord étudié le corpus POMELO précédemment construit dans le projet du même nom [107]. Ce corpus avait été constitué en effectuant la requête suivante dans MEDLINE (la précision "MH" entre crochets

11. `http://www.theriaque.org`

indique que le terme recherché devait être parmi les termes MeSH utilisés pour indexer les articles dans MEDLINE) :

```
(‘‘FOOD DRUG INTERACTIONS’’[MH] OR ‘‘FOOD DRUG INTERACTIONS*’’) AND (‘‘adverse effects*’’)
```

Cependant, une analyse bibliographique des références citées dans un manuel de référence sur les interactions entre les plantes médicinales et les médicaments, le Stockley (Stockley's Drug Interactions, 8th edition) [12], qui mentionne des interactions avec certains aliments ayant été jugées pertinentes par les experts, a révélé que le corpus POMELO ne couvrait que 3% des articles cités dans le Stockley [108]. Georgeta a alors constitué un autre corpus à partir du Stockley, contenant 1610 abstracts.

Réutilisation de RTOs existantes. Parallèlement, nous avons examiné les RTOs qui pouvaient être réutilisées pour représenter les entités d'intérêt. Pour cela, nous avons utilisé BioPortal car c'est le répertoire qui contient le plus de RTOs du domaine biomédical [205].

Nous avons commencé par rechercher les RTOs décrivant les interactions entre médicaments car les experts ont signalé qu'elles présentent de nombreuses similitudes avec les interactions aliment-médicament. Les principales RTOs décrivant les interactions entre médicaments sont DIO (Drug Interaction Ontology) [210], DINTO (Drug-Drug Interactions Ontology) [211] et DIDEO (Drug-drug Interaction and Drug-drug Interaction Evidence Ontology) [212], toutes trois étant alignées à BFO [207]. Dans [212], les concepteurs de DIDEO ont souligné que les principaux inconvénients de DIO sont qu'elle contient des incohérences et qu'elle ne précise pas en quoi des entités différentes, telles que le médicament, la substance chimique et la molécule, se distinguent. En ce qui concerne DINTO, les mêmes auteurs ont indiqué qu'il n'est pas possible d'y décrire les interactions *potentielles* entre médicaments alors que c'est ce qui est décrit le plus souvent dans la littérature. Sur la base de ces observations, nous avons choisi de réutiliser DIDEO comme fondement de FIDEO.

Ensuite, nous avons cherché si une RTO visant à décrire les aliments existait. Lors de l'analyse des interactions mentionnées dans les textes du corpus, nous avons constaté qu'en plus des aliments eux-mêmes et leurs catégories (*e.g.*, légumes crucifères, aliments contenant de la tyramine), il était également nécessaire de représenter leurs modes de cuisson, de préparation et/ou de conservation puisqu'ils peuvent être impliqués dans l'interaction (*e.g.*, viande grillée, thé infusé, jus de pamplemousse surgelé). En recherchant ces termes dans BioPortal, nous avons identifié qu'un concept Food existait dans ChEBI (Chemical Entities of Biological Interest) [213], ce qui avait un intérêt particulier car cette RTO est déjà utilisée dans DIDEO pour représenter les substances chimiques. Cependant, ce concept est défini comme un rôle selon BFO (role est un descendant de specifically dependent continuant) tandis que les substances chimiques (qui servent à décrire les médicaments) sont des entités matérielles (material entity est un descendant de independent continuant). Comme il est plus cohérent de représenter ces deux types d'objets de la même façon comme des entités matérielles, nous n'avons pas choisi ChEBI pour représenter les aliments. La RTO FoodOn (Food Ontology) [214] nous a semblé répondre aux différents besoins identifiés, avec une représentation des aliments comme entités matérielles et la présence des concepts Food transformation process, Food cooking process et Food preservation process.

12. https://www.pharmpress.com/product/9780857113474/stockley

Enfin, un aspect important concerne les types d'interaction. Au moment de soumettre le projet MIAM, nous n'en avions pas mesuré toute la complexité. Une illustration de celle-ci est la modification des paramètres pharmacocinétiques (*e.g.*, augmentation de la concentration plasmatique, diminution de la biodisponibilité) des médicaments due aux interactions. En étudiant ce qui était décrit dans DIDEO, les experts ont rapidement indiqué que cela ne convenait pas car les types d'interaction n'y apparaissent pas. En revanche, DINTO contient un concept DDI mechanism (Drug-Drug Interaction mechanism) et des sous-concepts pertinents, même si la couverture des types d'interaction reste incomplète. Nous travaillons actuellement sur cette partie afin d'enrichir la représentation des interactions de DINTO, notamment d'après INO (Interaction Network Ontology) [215] et les types d'interaction décrits dans la base de données Thériaque.

Conceptualisation

La figure 4.3 illustre la manière dont les concepts de haut niveau de FIDEO sont reliés entre eux et avec les concepts de BFO.

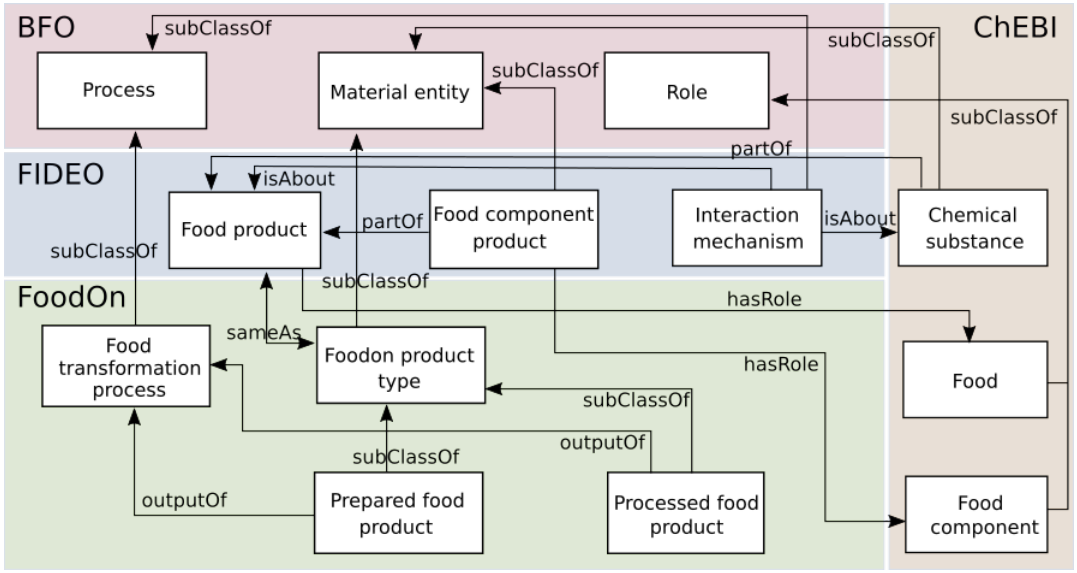

FIGURE 4.3 – Organisation des concepts de haut niveau de FIDEO selon la RTO dont ils sont issus.

Intégration

Cette étape a été réalisée manuellement afin d'articuler au mieux les concepts issus des RTOs que nous avons réutilisées. Ainsi, les médicaments (modélisés dans la figure 4.3 par le concept Chemical substance, sachant que le concept Drug product est relié à celui-ci dans DIDEO suivant une relation *partOf*) et les aliments sont des entités matérielles dans BFO. Par souci de simplification, nous avons introduit un concept Food product équivalent au concept Foodon product type. Celui-ci est relié au concept Food de ChEBI via une relation *hasRole*, qui modélise le fait que l'aliment en tant qu'entité matérielle ne peut interagir avec un médicament que s'il est effectivement utilisé dans son rôle nutritif (*i.e.*, consommé). Les composants alimentaires Food component product sont rattachés à Food product par une relation *partOf*. Le concept Chemical substance est relié à Food product par une relation *partOf* car certains aliments peuvent contenir des sub-

stances chimiques. Ce lien peut s'avérer utile pour inférer une interaction potentielle avec un aliment constitué d'une substance chimique interagissant avec un autre substance chimique dans le cadre d'une interaction entre médicaments. Le concept Food transformation process permet de décrire les processus de transformation que peuvent subir les aliments (via une relation *outputOf*). Enfin, le concept Interaction mechanism, concept DINTO renommé puisque sa hiérarchie va être enrichie et modifiée pour représenter les interactions entre un aliment et un médicament, est défini en tant que processus BFO, comme toute interaction dans INO. Les concepts Food product et Chemical substance sont reliés à Interaction mechanism par une relation isAbout.

Implémentation

L'implémentation des concepts de haut niveau de FIDEO a été réalisée manuellement à l'aide du logiciel Protégé [216], tout comme l'importation des différentes RTOs car nous avons décidé d'inclure uniquement les médicaments et les aliments impliqués dans au moins une interaction mentionnée dans notre corpus. Ce choix a été motivé par le fait que des RTOs telles que ChEBI et FoodOn contiennent de nombreux concepts non pertinents pour notre objectif. Jusqu'ici, nous avons utilisé 20 abstracts qui ont été annotés par des experts du CNHIM et du centre de pharmacovigilance de Bordeaux pour implémenter FIDEO. Pour trouver une correspondance à chaque entité annotée dans les RTOs que nous réutilisons, Georgeta a utilisé l'outil Stardog [13]. Celui-ci permet, entre autres, de stocker des RTOs et d'effectuer des requêtes SPARQL. Elle a également intégré des synonymes (récupérés dans l'UMLS) et normalisé les termes de manière basique (*e.g.*, suppression du pluriel, des caractères spéciaux, conversion en minuscules). Chaque entité annotée a donc été recherchée dans les RTOs chargées dans Stardog. Si une correspondance était trouvée, Georgeta a importé le concept et ses concepts descendants depuis la RTO appropriée et sinon, elle l'a ajouté à la main dans FIDEO.

Évaluation

FIDEO n'étant pas encore finalisée, l'étape d'évaluation en est à ses prémices. Suite à l'implémentation de FIDEO avec les 20 abstracts annotés par les experts, la couverture suivante a été obtenue pour les entités principales :

— 82% des médicaments et 90% des classes de médicaments ont été trouvés dans ChEBI,

— 62% des aliments et 76% des composants alimentaires ont été trouvés dans FoodOn.

Ces résultats encourageants indiquent que ces RTOs ont une portée acceptable pour la représentation des médicaments et des aliments impliqués dans des interactions.

Notons enfin que notre choix d'utiliser les entités annotées par les experts dans les abstracts garantit que le contenu de FIDEO couvre le domaine visé ; à condition de pouvoir l'enrichir avec les connaissances des autres abstracts du corpus et de futurs articles sur le sujet.

Documentation

L'article présentant FIDEO et la manière dont nous l'avons conçue est en cours de rédaction.

13. `https://www.stardog.com`

4.2.3 Conclusions

La conception de FIDEO relève des catégories de techniques linguistiques dites “seed words” et celles exploitant un lexique sémantique d’après la classification de Wong *et al.* [198]. En effet, nous utilisons le corpus de textes constitué à partir du Stockley pour identifier les termes (correspondant à cette notion de “seed words”) qu’il faut inclure dans FIDEO pour représenter les interactions décrites dans les abstracts en question. Par ailleurs, l’approche suivie est essentiellement basée sur l’intégration de RTOs existantes (jouant le rôle de lexique sémantique) que nous enrichissons afin de couvrir au mieux le domaine. Pour l’instant, de nombreuses étapes sont réalisées manuellement mais nous envisageons d’automatiser une partie de l’intégration et de l’implémentation. L’intégration peut être facilitée par des méthodes introduites au chapitre 3. En particulier, nous prévoyons d’utiliser une RTO de support décrivant les substances actives contenues dans les aliments afin d’établir le lien entre les aliments de FoodOn et les substances chimiques de ChEBI. Dans ce cadre, le travail décrivant la manière dont les RTOs GO et ChEBI ont été intégrées peut aussi constituer une bonne base de départ [142]. Côté implémentation, nous planifions d’automatiser l’importation grâce à l’outil Stardog (qui prend en entrée deux nœuds d’un arbre XML et retourne l’ensemble des nœuds intermédiaires).

FIDEO est une ontologie plus formelle qu’OntoAD puisqu’elle contient des concepts, des relations hiérarchiques et associatives ainsi que des axiomes. En effet, si l’on reprend l’exemple introduit dans la partie 4.2.1, la description d’une interaction entre la lévothyroxine et les aliments riches en fibres nécessite d’ajouter notamment l’axiome suivant aux aliments concernés : $\exists$ *contains*.Fiber. Cette description est cependant incomplète puisque ce qu’il faut représenter est en fait un aliment contenant une grande quantité de fibres. Sachant qu’il n’y a pas de précision quant au seuil permettant de déterminer qu’un aliment est *riche* en fibres, nous n’avons pour l’instant pas modélisé cette notion. Par ailleurs, il serait utile d’inclure des règles dans FIDEO si l’on souhaite pouvoir déduire qu’il existe une interaction potentielle avec un aliment donné s’il contient une substance chimique déjà impliquée dans au moins une interaction entre médicaments. Notons finalement que la composante temporelle liée au moment de la prise du médicament n’a pas encore été intégrée dans FIDEO.

Quand nous avons cherché des RTOs décrivant les aliments, nous avons remarqué que beaucoup de ressources existaient mais que peu d’entre elles étaient représentées de manière formelle. Une recherche dans BioPortal de RTOs dont le nom contient “food” illustre ce constat car seule FoodOn est dédiée exclusivement à la description des aliments parmi les 765 RTOs qui sont recensées dans ce portail. Cependant, d’autres RTOs comme le MeSH et le NCIt contiennent une hiérarchie des aliments, tandis que d’autres RTOs listent certains aliments jugés pertinents pour leurs propres besoins, comme la Stroke Ontology [14] qui décrit les légumes crucifères pour leur effet protecteur contre les attaques cérébrales. Une piste à approfondir est l’intégration des différentes connaissances liées aux aliments, y compris celles précisant leur impact (facteur protecteur ou aggravant) sur les maladies puisqu’une étude très récente a montré qu’une mort sur cinq était due à une mauvaise alimentation [217].

14. `http://bioportal.bioontology.org/ontologies/STO-DRAFT`

4.3 Synthèse et perspectives

De ce chapitre, nous pouvons dégager les **constats** suivants : (i) l'exploitation de ressources externes contribue à différentes tâches de la conception d'une RTO, (ii) une interaction forte avec les experts est nécessaire pour cette étape essentielle du cycle de vie d'une RTO et, (iii) les approches présentées aux chapitres 2 et 3 peuvent être utiles à l'activité de développement.

L'utilisation de RTOs existantes (ou lexiques sémantiques pour Wong *et al.* [198]) nous a servi pour former les concepts et leur attribuer un terme préféré, pour acquérir des relations hiérarchiques et associatives ainsi que pour obtenir des synonymes associés aux concepts. C'est donc un processus majeur des travaux que nous avons présentés dans ce chapitre. Dans une revue de la littérature, Simperl a identifié les quatre étapes suivantes comme étant nécessaires pour garantir une bonne réutilisation d'ontologies [191] :

1. l'identification des ontologies potentiellement pertinentes pour décrire la nouvelle ontologie,
2. la sélection des ontologies pouvant être réutilisées, et en particulier l'existence d'un serveur où sont stockées les ontologies jugées pertinentes,
3. la conversion des ontologies sélectionnées dans un format commun,
4. l'intégration des ontologies sélectionnées, ou d'une partie de celles-ci, dans une nouvelle ontologie. Cette étape est différente du processus d'intégration illustré dans la section 3.2 dont le but est de pouvoir rendre interopérables des ontologies utilisées conjointement [162].

L'approche suivie pour concevoir FIDEO est conforme à ces recommandations puisque : 1) pour choisir les RTOs pertinentes, nous avons exploité les textes décrivant les connaissances du domaine et consulté les experts, 2) pour sélectionner les RTOs, nous avons utilisé BioPortal, 3) ce portail a par ailleurs permis de disposer des RTOs sélectionnées au format OWL, et 4) FIDEO a été développée en intégrant une partie de ces RTOs.

Comme pour l'activité d'évaluation (chapitre 2), l'intervention d'experts du domaine pour lequel une RTO est créée est primordiale. Nous avons bénéficié de leur aide pour valider le contenu d'OntoAD, comme préconisé dans la méthodologie Terminae [206]. Lors de la conception de FIDEO, nous avons fait appel aux experts pour annoter les textes décrivant les interactions médicament-aliment ainsi que pour vérifier avec eux la pertinence des RTOs que nous avions identifiées pour représenter les différents types d'interactions, comme cela est recommandé par Simperl [191]. Notons finalement que des échanges sont indispensables entre les experts et la personne en charge de concevoir la RTO afin qu'elle s'assure que les connaissances décrites par les experts puissent être effectivement représentées dans la RTO (*e.g.*, selon la logique de description).

Lorsque nous avons créé OntoAD, l'utilisation de l'UMLS comme ressource externe a requis d'effectuer quelques corrections afin de disposer de connaissances cohérentes. Dans FIDEO, des adaptations sont faites quand les concepts issus des RTOs intégrées ne conviennent pas tout à fait à nos besoins. Lors du développement d'une RTO, il est donc important d'appliquer des méthodes d'évaluation de la qualité des RTOs que l'on réutilise [191], telles que celles présentées au chapitre 2. D'autre part, la conception d'une RTO suivant un processus d'intégration est facilitée si des techniques sont mises en œuvre pour identifier de manière automatique des correspondances entre les concepts et les relations des RTOs qui sont réutilisées. Les deux travaux présentés dans ce chapitre contribuent par ailleurs à augmenter le nombre déjà important de RTOs disponibles

en santé même si leur création était nécessaire de par la spécificité des connaissances qu'il fallait décrire. Ces constats illustrent l'importance des travaux menés sur l'interopérabilité entre RTOs abordés au chapitre 3. La capacité à utiliser conjointement des RTOs génériques de référence et des RTOs très spécialisées offre, d'un côté, une large couverture du domaine biomédical et garantit, d'un autre côté, une description très précise des connaissances.

Les **perspectives** des travaux développés dans cette section sont d'étudier des techniques différentes de celles que nous avons utilisées pour l'acquisition de connaissances. Au regard de la classification de Wong *et al.*, il reste des techniques linguistiques à investir, mais il faut surtout analyser ce que peuvent apporter les techniques statistiques et logiques. Dans leur revue de la littérature, les auteurs soulignent en particulier les efforts nécessaires pour faciliter l'acquisition des axiomes qui a fait l'objet de très peu de recherches [198]. Un autre champ qu'il faut continuer à approfondir est l'évaluation des RTOs créées, selon les deux facettes suivantes (en plus de la cohérence de leur contenu) : leur adéquation par rapport à l'application pour laquelle elles ont été conçues (*e.g.*, pour répondre à des questions telles que : est-ce que la recherche d'information résulte en de meilleurs résultats lorsque OntoAD est utilisée ?) et leur couverture du domaine (*e.g.*, est-ce que FIDEO contient l'ensemble des termes que les experts annotent dans le corpus issu du Stockley ?) [109]. Enfin, il serait intéressant de comparer le modèle de représentation d'OntoAD à d'autres modèles. Nous avons choisi de représenter les concepts comme des classes OWL et leurs termes comme des propriétés $skos{:}prefLabel$ et $skos{:}altLabel$ de ces classes. D'autres représentations, spécifiques aux RTOs, où les termes sont modélisés par des classes ont été proposées dans la littérature [218, 219, 220]. Une étude approfondie de ces modèles est nécessaire pour analyser en quoi ces représentations alternatives seraient bénéfiques à OntoAD afin de tirer le meilleur profit de sa forte composante terminologique, notamment au sein de SemBiP.

CHAPITRE 5

TRAVAUX EN COURS ET PERSPECTIVES

Dans ce chapitre, je présente tout d'abord les travaux que nous réalisons actuellement avec des collègues du LaBRI sur l'interprétation des données omiques grâce à la visualisation et aux RTOs ainsi que les perspectives offertes par la prise en compte des données cliniques dans ce cadre (section 5.1). Ensuite, j'introduis les pistes de recherche que j'envisage d'explorer dans les années à venir (section 5.2).

5.1 Interprétation de données omiques

Jusqu'à récemment, les données cliniques et les données omiques étaient traitées et exploitées de manière indépendante. Par exemple, la prise en charge des patients se faisait exclusivement à l'aide des données cliniques tandis que l'interprétation fonctionnelle des gènes se basait principalement sur des données et connaissances omiques. Avec le développement de la médecine personnalisée et des thérapies ciblées, les données omiques apportent des informations essentielles à la décision thérapeutique, au même titre que les données cliniques du patient [221]. Parallèlement, être en capacité de connaître le rôle des gènes dans les maladies ou encore dans la réponse d'un patient à un médicament est déterminant. De même, l'interprétation biologique d'un groupe de gènes au regard de son implication dans une ou des maladies est un défi auquel se confrontent des approches très diverses. L'émergence de ces besoins implique de s'intéresser à l'utilisation conjointe des données et des connaissances cliniques et omiques.

J'ai investi ces questions en développant des collaborations interdisciplinaires. Avec Patricia Thébault, maître de conférences en bioinformatique, nous encadrons un doctorant qui cherche à générer une annotation synthétique et pertinente de groupes de gènes (section 5.1.1) en exploitant des sources de connaissances omiques et cliniques (section 5.1.2). Parallèlement, nous avons mis en place un groupe de travail interdisciplinaire afin d'identifier les opportunités offertes par la visualisation pour répondre à ces problématiques. Notre première réalisation a été une revue de la littérature des méthodes existantes pour visualiser les données omiques et cliniques (section 5.1.3).

5.1.1 Annotation de groupes de gènes basée sur une ressource termino-ontologique

Notre doctorant, Aarón Ayllón-Benítez, a proposé une méthode d'annotation alternative aux approches d'enrichissement. Ces dernières sélectionnent les termes sur-représentés, c'est-à-dire ceux qui annotent une plus grande proportion des gènes se trouvant dans le groupe étudié par rapport à la proportion des gènes d'un génome donné. Le problème est que cela a tendance à privilégier l'annotation des gènes bien connus et à déprécier celle des gènes qui sont moins bien annotés [222, 223]. Par ailleurs, de nombreux outils d'enrichissement fournissent un grand nombre de termes, certains étant très proches sémantiquement et donc potentiellement redondants, sans pour autant couvrir l'ensemble des gènes du groupe étudié. Pour contourner ces limites, nous avons choisi d'utiliser la similarité sémantique afin de déterminer les termes proches sémantiquement et de fournir ainsi une annotation non redondante. De plus, pour que l'annotation finale soit synthétique avec un nombre restreint de termes couvrant un maximum de gènes du groupe, Aarón a développé un algorithme permettant d'identifier les termes d'annotation qui sont représentatifs d'un groupe de gènes.

Dans un premier article, nous avons étudié l'impact d'utiliser des mesures de similarité sémantique pour regrouper et synthétiser les informations pertinentes en biologie [91]. Pour cela, nous nous sommes intéressés à des mesures de similarité sémantique appartenant à des catégories différentes, telles que définies dans trois revues de la littérature [45, 46, 47]. Plus précisément, neuf mesures ont été sélectionnées : cinq mesures basées sur les termes de la Gene Ontology (GO) [36], trois mesures exploitant les liens entre ces termes et une mesure hybride. Pour que le panel étudié soit diversifié, nous avons choisi des mesures prenant en compte des caractéristiques différentes des termes et/ou liens entre eux (Figure 5.1).

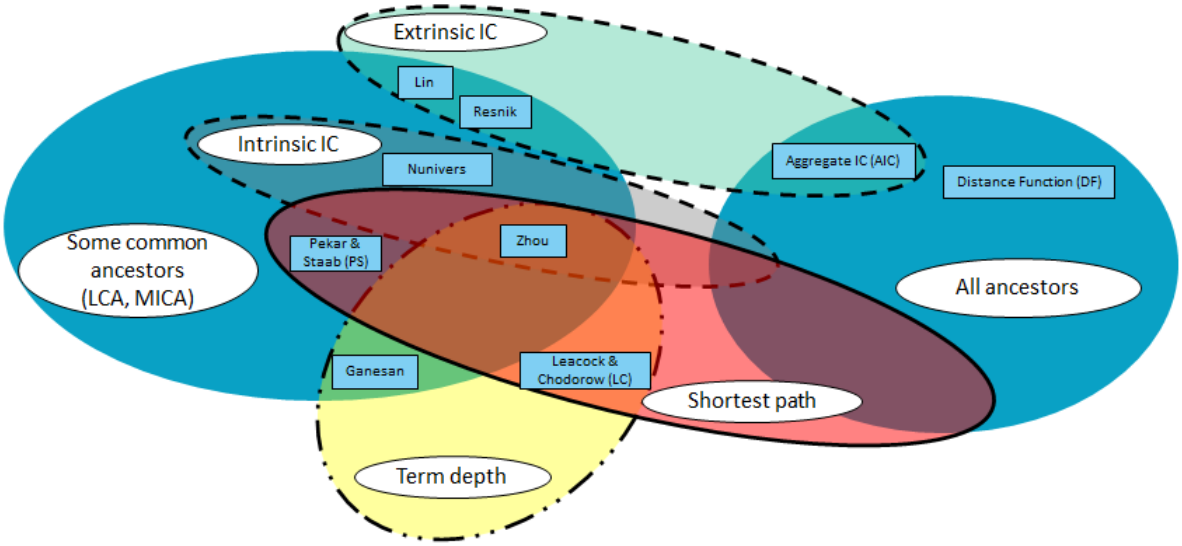

FIGURE 5.1 – Classification des neufs mesures de similarité sémantique investiguées dans Ayllón-Benítez *et al*, d'après les caractéristiques exploitées par les différentes formules (source : [91], adaptée de la figure 1 de [46]). IC (*information content*) correspond au contenu en information, LCA (*lowest common ancestor*) correspond à l'ancêtre commun le plus spécifique et MICA (*most informative common ancestor*) correspond à l'ancêtre commun ayant le meilleur IC. Les ICs intrinsèques prennent en compte uniquement les informations se trouvant dans la Gene Ontology tandis que les ICs extrinsèques exploitent le nombre de gènes annotés par les termes de la Gene Ontology.

Afin d'évaluer l'impact d'utiliser une mesure de similarité sémantique plutôt qu'une autre pour l'interprétation d'un groupe de gènes, Aarón a mis en œuvre une méthode d'annotation (détaillée au paragraphe suivant) que nous avons appliquée sur des données humaines d'immunologie. Nous avons vérifié si les résultats étaient les mêmes, quelle que soit la méthode de classification ascendante hiérarchique choisie (comparaison des résultats entre *single*, *complete* et *average* qui diffèrent dans leur manière d'intégrer les termes au moment de la constitution des classes, ou *clusters*). Ensuite, nous avons étudié la capacité de chaque mesure à générer une annotation pertinente et synthétique du groupe de gènes d'intérêt, qui soit représentative du plus grand nombre de ces gènes. Les meilleurs résultats ont été obtenus en utilisant les mesures de similarité sémantique basées sur les termes, quelles que soient les caractéristiques qu'elles exploitent.

Suite à ce travail, Aarón a créé GSAn [1] (**G**ene **S**et **An**notation) qui est une application Web permettant de visualiser l'annotation unifiée et synthétique d'un groupe de gènes d'un organisme donné [92]. La méthode qu'il a mise en œuvre consiste en quatre étapes successives (Figure 5.2) :

1. Suppression des termes d'annotation inappropriés. Parmi les termes GO associés à chaque gène du groupe récupérés dans le fichier Gene Ontology Annotation [2] de l'organisme étudié, sont éliminés les termes d'annotation redondants et incomplets. Ainsi, si deux termes reliés hiérarchiquement annotent un même gène, le terme le plus générique est considéré comme redondant car il apporte une annotation moins précise que l'autre terme. Au sens de Faria *et al.*, les annotations incomplètes sont peu informatives [224]. Nous considérons que ce sont celles impliquant les termes GO qui ont un contenu en information relativement bas (*i.e.*, inférieur au premier quartile du contenu en information de l'ensemble des termes GO).
2. Classification hiérarchique des termes d'annotation. Une matrice de similarité sémantique entre les termes d'annotation restants est créée à partir d'une des cinq mesures basées sur les termes identifiées comme les meilleures dans [91]. La méthode de classification *average* est utilisée pour constituer une partition des termes d'annotation et générer les clusters de termes (car c'est celle qui a permis d'obtenir les meilleurs clusters).
3. Identification des termes **représentatifs**. Les termes d'un même cluster présentant un certain degré de similarité, il est *a priori* possible de sélectionner seulement certains d'entre eux (ou une combinaison de leurs parents et/ou ancêtres) pour représenter les gènes annotés par ces termes. L'algorithme développé par Aarón pour déterminer les termes représentatifs d'un cluster est décrit dans [91]. Une fois ces termes identifiés pour chaque cluster, ils sont regroupés et trois filtres sont appliqués afin d'éliminer les termes représentatifs : 1) ayant un contenu en information trop bas, 2) annotant moins de trois gènes du groupe d'intérêt, et 3) étant parents ou ancêtres d'un autre terme représentatif.
4. Sélection des termes **synthétiques**. Pour générer un nombre limité de termes, qualifiés de synthétiques, Aarón a implémenté un algorithme relevant du problème de couverture par ensembles. Pour un ensemble E d'éléments, il calcule la plus petite combinaison de sous-ensembles de E couvrant tous ses éléments. Ici, chaque gène est un élément et chaque terme représentatif est un ensemble d'éléments, correspondant aux gènes que ce terme annote. Pour un ensemble de termes représentatifs, on cherche donc la plus petite combinaison de termes représentatifs qui couvrant le maximum de gènes du groupe étudié.

1. `https://gsan.labri.fr`
2. `http://geneontology.org/page/download-go-annotations`

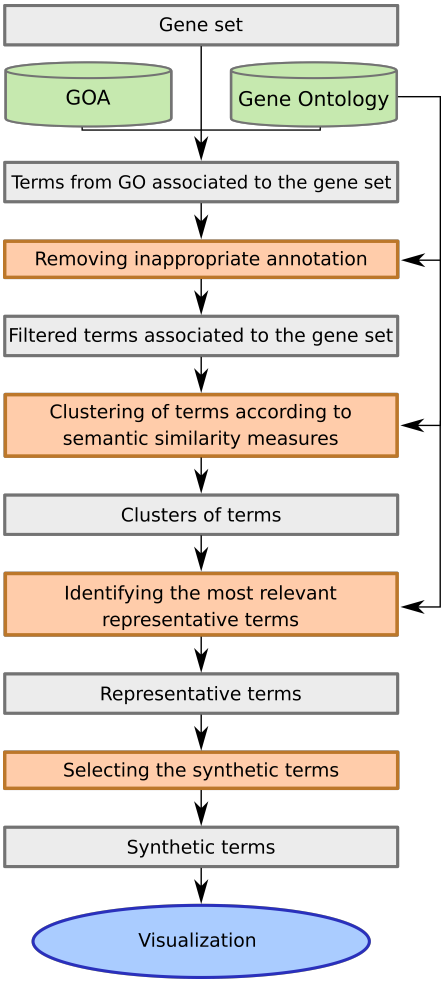

FIGURE 5.2 – Méthode implémentée dans GSAn (source : [92]). Les rectangles orange correspondent aux étapes de la méthode, les rectangles gris sont les objets manipulés en entrée et/ou en sortie des différentes étapes et les éléments en vert sont les ressources externes utilisées au cours de certaines étapes.

Pour faciliter l'exploration des résultats générés par GSAn, Aarón a élaboré une visualisation basée sur les relations hiérarchiques de GO. On peut y visualiser de manière interactive les termes représentatifs, leurs ancêtres et les gènes qu'ils annotent [93]. Celle-ci comprend deux métaphores visuelles (Figure 5.3) : (i) un arbre indenté, et (ii) un diagramme de répartition circulaire (*circular treemap* introduit initialement par Kai Wetzer [3]). Cette visualisation nécessitant de disposer d'un arbre, seul le parent ayant le meilleur IC a été sélectionné pour chaque terme représentatif. Par ailleurs, Aarón a adapté un algorithme permettant d'attribuer des couleurs similaires à des termes reliés hiérarchiquement. Dans le diagramme de répartition circulaire, chaque cercle correspond à un terme GO, sauf les feuilles qui sont représentées par des cercles blancs et correspondent aux gènes annotés par le terme dans lequel ils se trouvent. Dans chacun de ces cercles blancs est présenté un diagramme en barre des termes annotant ce gène. En termes d'interaction, un clic sur le nom des termes ou des gènes développe l'arborescence au sein de l'arbre indenté et permet de visualiser l'élément correspondant dans le diagramme de répartition circulaire. Un clic sur les cercles a aussi pour effet de développer l'arborescence de l'arbre indenté et permet de zoomer sur le contenu du terme ou gène sélectionné.

3. `http://lip.sourceforge.net/ctreemap.html`

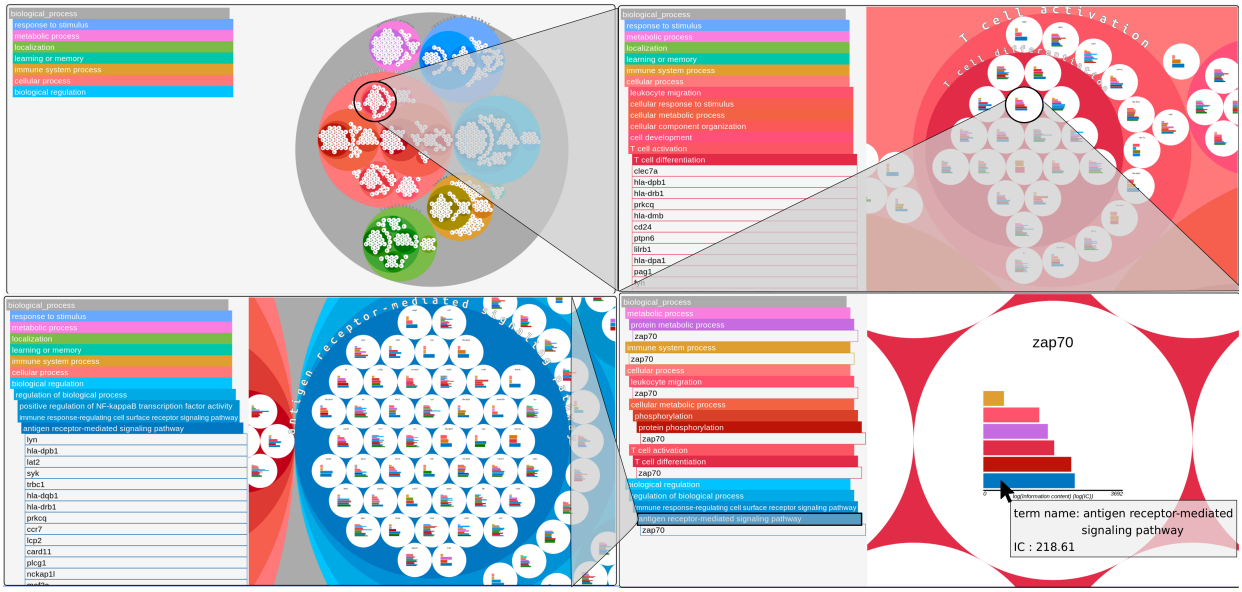

FIGURE 5.3 – Visualisation implémentée dans GSAn illustrant les deux métaphores mises en œuvre (source : [92]) : l'arbre indenté à gauche et le diagramme de répartition circulaire à droite, et les fonctionnalités d'interaction matérialisées par les triangles gris transparents : le zoom et le clic.

5.1.2 Annotation issue de multiples ressources termino-ontologiques

La dernière partie de la thèse d'Aarón a pour objectif d'utiliser d'autres RTOs que la GO pour annoter les gènes afin d'apporter des informations complémentaires, en particulier des connaissances cliniques, concernant les fonctions d'un groupe de gènes. Pour cela, Aarón étudie la possibilité d'intégrer deux RTOs supplémentaires dans GSAn : Reactome qui fournit des informations sur les voies métaboliques [225], et Disease Ontology qui décrit les maladies humaines [226]. L'intégration des termes d'annotation issus de multiples RTOs est classiquement réalisée *a posteriori* pour comparer les gènes (ce que fait par exemple le *clustering* effectué par l'outil d'enrichissement DAVID [227]). Cette étape est en cours de développement mais nous avons pour objectif d'intégrer les termes issus de multiples RTOs en amont afin d'effectuer cette intégration une seule fois, indépendamment des gènes à annoter (et non à chaque fois qu'un groupe de gènes est annoté), et d'affecter un poids aux termes d'annotation, qui pourra être déterminé en fonction du nombre de RTOs le spécifiant et du niveau de confiance accordé à chaque RTO. Dans ce cadre, les travaux autour du *Linked Open Data* présentent un intérêt certain [228, 229, 230]. Cependant, dans ces articles, les liens entre les différentes RTOs ne sont pas décrits au niveau conceptuel puisque ce sont des références croisées entre identifiants de gènes qui permettent de récupérer des informations issues de ces différentes RTOs, sans qu'il y ait forcément d'intégration au niveau des termes d'annotation eux-mêmes. En revanche, Callahan *et al.* ont exploité une ontologie comme pivot pour faire correspondre les connaissances de plusieurs RTOs intégrées dans Bio2RDF [231]. L'approche suivie dans ce travail est particulièrement intéressante même si les auteurs exploitent essentiellement des RTOs biologiques. Nos développements futurs pourront s'inspirer de ce travail en élargissant son champ avec des RTOs médicales.

5.1.3 Visualisation des données omiques et cliniques

Dans le cadre de nos échanges avec Patricia, nous nous sommes intéressées aux spécificités des données cliniques et omiques afin de comprendre les raisons pour lesquelles elles sont peu exploitées de manière conjointe. L'angle d'attaque que nous avons choisi est celui de la visualisation puisqu'elle replace l'expert au centre du processus d'analyse. Nous avons mis en place un groupe de travail interdisciplinaire réunissant des collègues enseignants-chercheurs spécialistes en analyse visuelle (*visual analytics*), en représentation des connaissances, en biostatistiques et en bioinformatique. Notre premier objectif a été de réaliser, de manière collaborative, une revue de la littérature des approches de visualisation d'information et des logiciels permettant de visualiser des données cliniques et omiques [90] afin d'identifier des pistes de recherche à investir, dans un deuxième temps. Après avoir illustré la variété des données cliniques et omiques à manipuler, nous avons décrit les métaphores visuelles existantes et précisé lesquelles sont privilégiées en fonction du type de données à visualiser. Ensuite, nous avons examiné les outils de visualisation des données omiques indépendamment de ceux offrant la possibilité d'analyser les données cliniques au regard du mantra de Shneiderman [232] : *Overview first, zoom and filter, then details-on-demand*. Cet auteur a défini ces recommandations pour faciliter l'exploration d'information lorsque la taille et la complexité des données rendent la tâche manuelle trop difficile. L'idée est de fournir des capacités d'interaction au sein des outils de visualisation pour que les utilisateurs puissent : (i) avoir une vue globale sur leurs données, (ii) filtrer, et (iii) zoomer sur les données afin de se focaliser sur une partie d'entre elles et les étudier plus en détails. Nous avons tout d'abord noté qu'il existait beaucoup plus d'outils pour visualiser les données omiques que pour les données cliniques. Par ailleurs, nous avons remarqué que les logiciels de visualisation des données omiques fournissent de nombreuses fonctionnalités d'interaction et un large panel de métaphores visuelles dont les plus populaires sont les diagrammes nœuds-liens, les graphiques de coordonnées parallèles ou encore les cartes thermiques (*heatmap*). En revanche, en ce qui concerne les données cliniques, la métaphore visuelle dominante est la représentation graphique chronologique (*timeline*) lorsque la visualisation doit être faite à l'échelle d'un patient donné tandis que des métaphores variées sont utilisées quand la visualisation de plusieurs patients est nécessaire. Nous nous sommes finalement intéressés aux logiciels qui permettaient l'exploration de données cliniques et omiques de manière simultanée (ceux-ci étant recensés dans [233]) et nous avons constaté qu'il en existait très peu. Un seul logiciel offre une visualisation des données temporelles [234], qui est pourtant une dimension essentielle du point de vue clinique. Ce premier travail nous a permis d'identifier qu'il y a un fort besoin d'intégration des données cliniques et omiques pour pouvoir les visualiser conjointement de manière efficace.

En plus des collaborations sur la visualisation de données omiques et cliniques, nous envisageons d'étudier les approches existantes pour visualiser des RTOs. Parmi les défis actuels mentionnés dans une revue récente de la littérature sur ce sujet [235], on peut citer le besoin de visualiser les larges RTOs de manière interactive ainsi que la nécessité d'adapter les méthodes et logiciels proposés à un cas d'utilisation réel. Les RTOs volumineuses du domaine biomédical présentent donc un intérêt particulier dans ce cadre. Par ailleurs, étudier l'intérêt non négligeable de bons "outils" de visualisation pour faciliter le travail des experts et des concepteurs au cours des différentes étapes du cycle de vie d'une RTO est, à mon sens, un cas d'usage prometteur [236, 237].

5.2 Perspectives

Dans cette section, j'aborde les pistes de recherche que je souhaite explorer dans les années à venir, au-delà des perspectives évoquées dans les chapitres précédents. En pratique, cela concerne la gestion de l'évolution des RTOs, l'exploitation des instances présentes dans les RTOs ainsi que les apports réciproques des RTOs et des méthodes d'apprentissage.

5.2.1 Évolution des ressources termino-ontologiques

Parmi les étapes du cycle de vie d'une RTO, je veux approfondir les enjeux liés à la gestion de son évolution. Cette étape est majeure, comme l'illustre la figure 5.4 qui représente le cycle de vie d'une base de connaissances (type de système de représentation qui peut être englobé dans le terme RTO) comme un processus itératif au sein duquel l'évolution joue un rôle essentiel [238].

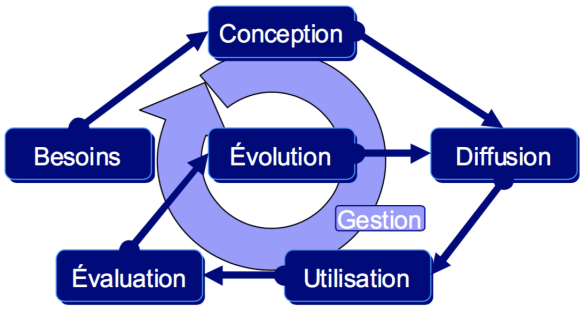

FIGURE 5.4 – Cycle de vie d'une base de connaissances décrit par Dieng *et al.* (source : [238]).

Dans le domaine de la santé, cette étape est d'autant plus importante que le contenu des RTOs biomédicales évolue constamment [239] puisque la moitié des connaissances médicales se périment au bout d'environ cinq ans [240]. Une illustration de ce constat est l'évolution des connaissances sur les interactions médicament-aliment. Au cours du projet MIAM, Georgeta a comparé manuellement la version du Stockley's Herbal Medicines Interactions de 2008 ($8^{\text{ème}}$ édition) avec celle de 2016 ($11^{\text{ème}}$ édition). Sur 269 interactions listées dans la version de 2008, 70 d'entre elles n'apparaissent plus dans la version de 2016 et 113 nouvelles interactions ont été décrites. Des explications possibles à ces différences sont les suivantes : (i) l'interaction a été jugée non significative entre temps, (ii) l'interaction a été généralisée ou précisée (*i.e.*, décrite initialement avec un aliment/médicament puis généralisation à une catégorie plus générale d'aliments/de médicaments, ou inversement), (iii) le médicament n'est plus commercialisé. La principale difficulté réside dans la manière d'identifier ces évolutions de manière automatique, comme par exemple de détecter quels articles ont mené à ces modifications dans le Stockley. Un autre aspect lié à cette problématique dans le cadre du projet MIAM est la gestion de l'évolution des RTOs sur lesquelles repose FIDEO.

Flouris *et al.* ont déclaré que les problématiques liées à l'évolution des RTOs et à l'évaluation de leur qualité étaient très similaires [111]. Dans le premier cas, il faut gérer l'ajout de nouvelles connaissances, la modification et la suppression de connaissances existantes en vérifiant que cela ne génère pas de contradictions dans la RTO concernée. Ce processus est donc assez proche de l'évaluation de la qualité des RTOs qui vise à identifier, et parfois à résoudre, d'éventuelles incohérences. Duque-Ramos *et al.* ont d'ailleurs montré que le cadre global d'évaluation de la

qualité des RTOs qu'ils ont défini, OQuaRE, a pu être adapté facilement pour analyser les différences entre plusieurs versions d'une RTO et identifier celles qui ont fait l'objet de modifications conséquentes [241]. Les travaux présentés au chapitre 2 peuvent donc contribuer à gérer l'évolution d'une RTO en cas de modifications de ses connaissances. Cependant, il faut aussi veiller à maintenir la cohérence de la RTO au regard des utilisations qui en sont faites, par exemple lorsque la RTO sert à annoter des documents [242] et quand ses entités sont liées aux entités d'autres RTOs [239]. L'option la plus simple consiste à supprimer l'existant et à tout re-générer avec les nouvelles connaissances, ce qui ne convient pas en santé où l'historisation de l'évolution des connaissances est indispensable. D'autres approches visant à identifier automatiquement les changements ayant eu lieu et à proposer aux experts les évolutions à faire dans la RTO sont donc mieux adaptées au domaine biomédical.

La gestion des versions d'une RTO est parfois distinguée de l'évolution dans la littérature. Ce processus est particulièrement complexe lorsqu'il implique de maintenir la validité de l'ensemble des versions qui doivent co-exister et de garantir leur interopérabilité [111]. Je compte investir cette problématique importante au travers d'une application concrète en santé : le passage de la CIM-10 à la CIM-11. La CIM dont la 11$^{\text{ème}}$ édition est désormais disponible est un exemple de RTOs dont la gestion de l'évolution est majeure. Actuellement, la CIM-10 est la version la plus largement utilisée pour le codage des diagnostics dans les hôpitaux et des causes de décès. La CIM-11, version suivante, a pour objectifs ambitieux d'élargir le champ des connaissances représentées, d'assurer son lien avec des RTOs de référence et d'être décrite de manière formelle. Elle repose sur une approche collaborative de mise à jour via une plateforme Web où les experts peuvent l'enrichir et la modifier [243]. Accessible depuis 2010, cette plateforme a déjà eu un large succès avec l'intervention de 270 experts du domaine en l'espace de trois ans [244]. Malgré ces caractéristiques très riches et prometteuses du point de vue informatique, l'adoption de la CIM-11 par la communauté médicale n'est pas acquise. L'expérience du passage de la CIM-9-CM à la CIM-10-CM en 2015 aux États-Unis ayant déjà fait l'objet de nombreux débats [245], l'acceptation d'un nouveau changement n'en sera pas facilité. L'utilisation de la CIM-11 étant prévue pour 2022 par l'OMS (organisation mondiale de la santé), cette transition complexe nécessite la participation active de la communauté d'informatique médicale.

Les deux RTOs sont très différentes. De nouveaux chapitres ont été introduits dans la CIM-11, des chapitres de la CIM-10 ont été réorganisés et il est possible d'utiliser le mécanisme de post-coordination pour coder avec la CIM-11. Grâce à ce mécanisme, les praticiens pourront combiner des diagnostics avec d'autres types de concepts, par exemple pour préciser la sévérité d'une maladie ou la latéralité de l'organe concerné. Parmi les problématiques liées à l'interopérabilité qui se poseront quand la CIM-11 sera effectivement utilisée, je vais m'attaquer aux suivantes :

— l'OMS prévoit de fournir l'alignement entre la CIM-10 et la CIM-11 mais sans mettre à disposition de mappings vers les autres RTOs de référence telles que la SNOMED CT,

— grâce aux nouveaux types de concepts décrits dans la CIM-11 pour élargir le champ couvert, des correspondances pourront être recherchées avec des RTOs représentant d'autres connaissances que les diagnostics. Par exemple, la classe "Substances" contient des concepts également décrits dans ChEBI,

— la CIM-11 est définie dans le langage formel OWL mais, à ma connaissance, elle n'a été pas rattachée à une ontologie de haut niveau telle que BFO.

Les techniques proposées au chapitre 3 permettront d'aborder en partie ces questions.

5.2.2 Intérêt des instances dans les ressources termino-ontologiques

Une autre piste de recherche que j'envisage concerne l'exploitation des instances au sein des RTOs. C'est d'autant plus intéressant d'explorer cet aspect que la plupart des RTOs biomédicales ne contiennent pas d'instances. Une première étape pour étudier cette question sera donc d'instancier les RTOs étudiées afin de disposer de ce type d'entités.

Les instances sont communément utilisées pour établir des correspondances entre les concepts auxquels elles sont rattachées. Les techniques correspondantes sont dites *instance-based* dans la classification d'Euzenat et Shvaiko [165] et peu de systèmes participant aux campagnes OAEI les mettent en œuvre, en comparaison avec les techniques terminologiques et structurelles [164]. J'envisage d'étendre le travail d'Aarón en mettant à profit mes travaux de thèse où j'avais proposé une méthode d'alignement basée sur les instances [68, 70]. Une piste intéressante pour établir des liens entre RTOs est de se servir des gènes qui sont annotés par leurs concepts, les gènes jouant ainsi le rôle d'instances. Par exemple, les concepts de GO et de Disease Ontology annotant un même ensemble de gènes peuvent être mis en correspondance pour générer de nouvelles connaissances sur les liens entre gènes et maladies. Une approche alternative à investiguer pour relier les concepts de GO et de Disease Ontology est l'utilisation de méthodes d'apprentissage, comme cela a été proposé dans SAMBO [190]. Le principe est d'associer à chaque concept un corpus d'articles issus de MEDLINE où il apparaît et de générer un classifieur à partir de ces articles pour chaque RTO. Les articles d'une RTO sont ensuite classés par le classifieur de l'autre RTO (et donc rattachés aux concepts de cette RTO) et inversement, pour déterminer finalement la similarité entre les concepts des deux RTOs en fonction des articles qu'ils partagent. Quelle que soit la méthode implémentée, une question reste néanmoins entière : comment caractériser le lien entre les concepts qui ont été mis en correspondance ? Si la relation n'est ni équivalente, ni hiérarchique, une tâche importante et délicate consiste à la préciser.

Les éventuels bénéfices offerts par l'utilisation des instances des RTOs ont été globalement peu analysés. Convaincue de leur intérêt pour faciliter d'autres étapes du cycle de vie d'une RTO, je souhaite en étudier les possibilités. En particulier, leur utilisation par les raisonneurs peut être bénéfique pour faire des déductions supplémentaires à celles qui sont établies à partir des concepts. Typiquement, le projet MIAM en illustre l'intérêt et, plus précisément, l'instanciation possible de FIDEO à partir de produits alimentaires vendus dans le commerce via la base de données OpenFoodFacts [4]. Ces connaissances supplémentaires peuvent être exploitées pour déduire qu'il existe une interaction entre un produit alimentaire donné et un médicament (ou un classe de médicaments), ce qui est très utile en terme de santé publique. La base de données OpenFoodFacts et les applications mobiles basées sur celle-ci bénéficieraient de ces informations. Cette instanciation nécessite en premier lieu d'aligner les catégories auxquelles appartiennent les produits alimentaires définies dans OpenFoodFacts avec les aliments et les catégories d'aliments de FIDEO.

En outre, un travail a montré l'utilité des instances pour évaluer la couverture du domaine d'une RTO en la comparant à une autre RTO [246]. Une condition préalable est que les RTOs doivent partager le même ensemble d'instances, ce qui limite l'applicabilité de cette approche. Un autre article présente une méthode d'apprentissage basée sur les instances d'une RTO pour en évaluer la qualité [247]. Le principe consiste à estimer la pertinence d'un concept en fonction des

4. `https://fr.openfoodfacts.org`

instances qui lui sont spécifiques par rapport à celles de ses concepts voisins hiérarchiques (*i.e.*, ses parents et les enfants de ses parents). Les auteurs ont appliqué leur méthode sur la hiérarchie des composants cellulaires de la GO en l'instanciant avec des protéines mais les résultats n'ont pas été concluants car trop peu de concepts étaient instanciés. Ces travaux, bien que présentant des limites, constituent une base de réflexion.

5.2.3 Liens entre méthodes d'apprentissage et ressources termino-ontologiques

La place grandissante occupée par les méthodes d'apprentissage m'incite à considérer leurs apports potentiels aux RTOs, et inversement. Le fait que les RTOs du domaine biomédical soient volumineuses est d'ailleurs une opportunité pour cette catégorie de méthodes.

Des méthodes d'apprentissage ont été proposées pour le développement, la maintenance et l'alignement de RTOs. Elles ont été utilisées pour réaliser diverses tâches d'acquisition de connaissances à partir de textes, comme illustré dans la classification de Wong *et al.* [198]. Par exemple, la découverte de synonymes et la construction de hiérarchies peuvent être facilitées par des méthodes permettant de calculer la proximité entre deux termes d'après leur vecteur de termes cooccurant dans les textes d'un corpus donné. Lorsqu'elles sont basées sur des documents textuels, les étapes de développement et de maintenance de RTOs bénéficient naturellement de ces méthodes. L'utilisation de méthodes d'apprentissage a également contribué à la mise en correspondance de RTOs [165]. Les algorithmes sont entrainés à partir de mappings existants (corrects et incorrects) qui sont ensuite utilisés pour déduire de nouveaux mappings. Les types d'informations exploitées (*e.g.*, la fréquence des mots, les instances, les attributs) par les algorithmes varient en fonction des méthodes d'apprentissage implémentées. De par leur besoin d'apprendre sur un ensemble de mappings pré-existants, ces méthodes sont particulièrement adaptées aux situations où une RTO doit être mise en correspondance avec un ensemble de RTOs déjà alignées entre elles.

Récemment, l'intérêt de ces méthodes pour évaluer la qualité d'une RTO biomédicale a été étudié [248]. Les auteurs ont cherché à identifier des relations de subsomption manquantes dans le NCIt. Pour cela, ils ont testé deux algorithmes d'apprentissage supervisé. La plus grande difficulté a été de constituer l'ensemble d'exemples (*i.e.*, paires de concepts) négatifs nécessitant qu'aucune relation de subsomption n'existe entre les deux concepts, mais qu'ils soient proches malgré tout. De plus, les ensembles devant être équilibrés et le nombre d'exemples positifs étant fixe, il fallait sélectionner à peu près le même nombre d'exemples négatifs alors qu'il y en avait bien plus. Malgré ces difficultés et des résultats décevants, ce travail ouvre de nouvelles perspectives sur les possibles contributions des méthodes d'apprentissage au processus d'évaluation des RTOs.

L'autre enjeu tout aussi important est de déterminer si les performances des méthodes d'apprentissage peuvent être améliorées par la prise en compte de connaissances décrites dans les RTOs. Dans ce cadre, Ebert-Uphoff et Gil ont indiqué que le contenu d'une RTO peut être utilisé [249] :

— comme connaissances de base à fournir en entrée du modèle,

— pour sélectionner les variables d'entrée,

— pour contraindre le processus d'apprentissage,

— comme référence pour vérifier ou expliquer les résultats obtenus par des méthodes d'apprentissage.

Une illustration est le modèle MeSH-gram qui adapte l'algorithme *word2vec* [250] en remplaçant les mots du vecteur, à savoir les mots cooccurrents et les mots décrivant le contexte d'un mot donné dans un corpus d'articles scientifiques de MEDLINE, par les concepts MeSH utilisés pour indexer les abstracts où le mot en question apparaît [251]. Cet ajustement permet de disposer de vecteurs contenant les concepts qui sont *a priori* les plus pertinents pour décrire un article, tout en résolvant les limites liées à la synonymie et en réduisant le nombre de termes présents dans les vecteurs. Je souhaite approfondir ce type d'approches afin de tirer le meilleur profit des connaissances décrites dans les RTOs pour perfectionner les méthodes d'apprentissage.

CHAPITRE 6

CONCLUSION

Les trois activités du cycle de vie d'une RTO que j'ai abordées dans ce manuscrit s'alimentent les unes les autres. Le développement d'une RTO nécessite souvent de réutiliser d'autres RTOs, parfois via un processus d'intégration. De plus, les processus rendant possible l'utilisation conjointe de RTOs permettent indirectement d'évaluer la qualité des RTOs en question, mais aussi la RTO servant de support le cas échéant. Parallèlement, des besoins en terme d'évaluation émergent lors de la réutilisation de RTOs une fois l'intégration faite, puisque ce processus peut générer des erreurs. Mes travaux reflètent l'interrelation existant entre les différentes activités du cycle de vie d'une RTO et soulignent la nécessité de les réaliser de manière itérative, comme l'ont souligné Sure *et al.* [62].

La représentation des connaissances du domaine biomédical est une problématique de recherche à part entière, de par la complexité du langage médical et la très rapide évolution des connaissances en santé. Une conséquence de ces constats est que les RTOs disponibles dans ce domaine sont d'une grande diversité, qu'elles sont nombreuses, volumineuses et changeantes. Ces multiples caractéristiques illustrent l'intérêt d'étudier ce domaine d'application en particulier et d'effectuer des recherches dans le champ de l'informatique médicale. Shortliffe a souligné que, parmi les défis actuels de l'intelligence artificielle en santé, il y a un réel besoin que des chercheurs soient capables de prendre en compte l'interdisciplinarité de l'informatique médicale [252]. Ce n'est qu'en acquérant de solides connaissances en santé et en tissant une étroite collaboration avec les experts médicaux que l'on peut répondre à ses défis. C'est dans cette optique que s'inscrivent mes activités en tant qu'enseignante-chercheuse ; par la conduite de recherches en informatique médicale, l'encadrement de doctorants de plusieurs disciplines (Jean Noël est médecin de santé publique spécialisé en informatique médicale et Aaron a étudié la bioinformatique) et la formation d'étudiants de profils différents dans le cadre du Master 2 SITIS (systèmes d'informations et technologies informatiques pour la santé) dont je suis responsable depuis 2016.

Bibliographie

[1] Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. Journal of Global Health, 2:e020303, December 2018.

[2] Koichiro Yasaka and Osamu Abe. Deep learning and artificial intelligence in radiology: current applications and future directions. PLoS Medicine, 11:e1002707, November 2018.

[3] Kipp W. Johnson, Jessica Torres Soto, Benjamin S. Glicksberg, Khader Shameer, Riccardo Miotto, Mohsin Ali, Euan Ashley, and Joel T. Dudley. Artificial intelligence in cardiology. Journal of the American College of Cardiology, 71(23):2668–2679, 2018.

[4] Alex Zhavoronkov, Polina Mamoshina, Quentin Vanhaelen, Morten Scheibye-Knudsen, Alexey Moskalev, and Alex Aliper. Artificial intelligence for aging and longevity research: Recent advances and perspectives. Ageing Research Reviews, 49:49 – 66, 2019.

[5] Douglas Laney. 3D data management: Controlling data volume, velocity, and variety. Technical report, META Group, February 2001.

[6] Rob Kitchin and Gavin McArdle. What makes big data, big data? Exploring the ontological characteristics of 26 datasets. Big Data & Society, 3(1):2053951716631130, 2016.

[7] Stuart Russell. Unifying logic and probability: A new dawn for AI? In Information Processing and Management of Uncertainty in Knowledge-Based Systems, pages 10–14, Cham, 2014. Springer International Publishing.

[8] John McCarthy. Programs with common sense. In Proceedings of the Teddington Conference on the Mechanization of Thought Processes, pages 756–791. H.M. Stationery Office, 1959.

[9] Yue-ting Zhuang, Fei Wu, Chun Chen, and Yun-he Pan. Challenges and opportunities: from big data to knowledge in AI 2.0. Frontiers of Information Technology & Electronic Engineering, 18(1):3–14, Jan 2017.

[10] Daniel E. O'Leary. Artificial intelligence and big data. IEEE Intelligent Systems, 28(2):96–99, March 2013.

[11] Randall Davis, Howard E. Shrobe, and Peter Szolovits. What is a knowledge representation? AI Magazine, 14:17–33, 1993.

[12] Nathalie Aussenac-Gilles, Jean Charlet, and Chantal Reynaud-Delaître. Ingénierie des connaissances. In Représentation des connaissances et formalisation des raisonnements en I.A., volume 1 of Panorama de l'Intelligence Artificielle, chapter 1.20, pages 500–537. Cépaduès Editions, janvier 2014.

[13] Pierre Zweigenbaum. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. Innovation Stratégique en Information de Santé, pages 27–47, 1999.

[14] Daniel Delattre. Galien, Systématisation de la médecine. Presses du Septentrion, 2003.

[15] Melissa A. Haendel, Julie A. McMurry, Rose Relevo, Christopher J. Mungall, Peter N. Robinson, and Christopher G. Chute. A census of disease ontologies. Annual Review of Biomedical Data Science, 1(1):305–331, 2018.

[16] Josef Ingenerf and W Giere. Concept oriented standardization and statistics oriented classification: continuing the classification versus nomenclature controversy. Methods of Information in Medicine, 37:527–39, 1998 Nov 1998.

[17] C.K. Ogden and Ivor A. Richards. The Meaning of Meaning: a Study of the Influence of Language Upon Thought and of the Science of Symbolism. Harcourt Brace Jovanovich, 1923.

[18] Nicolette F. de Keizer, Ameen Abu-Hanna, and Johanna H. M. Zwetsloot-Schonk. Understanding terminological systems I: terminology and typology. Methods of Information in Medicine, 39(1):16–21, March 2000.

[19] Thomas R. Gruber. A translation approach to portable ontology specifications. Knowledge Acquisition, 5(2):199–220, June 1993.

[20] Rudi Studer, V. Richard Benjamins, and Dieter Fensel. Knowledge engineering: principles and methods. Data Knowledge Engineering, 25(1-2):161–197, March 1998.

[21] Ronald Cornet and Christopher G. Chute. Health concept and knowledge management: twenty-five years of evolution. Yearbook of Medical Informatics, Suppl 1:S32–41, 2016.

[22] Melissa A. Haendel, Christopher G. Chute, and Peter N. Robinson. Classification, ontology, and precision medicine. New England Journal of Medicine, 379(15):1452–1462, 2018.

[23] Songmao Zhang and Olivier Bodenreider. Comparing associative relationships among equivalent concepts across ontologies. Studies in Health Technology and Informatics, 107(Pt 1):459–466, 2004.

[24] Marie Chagnoux, Nathalie Hernandez, and Nathalie Aussenac-Gilles. An interactive pattern based approach for extracting non-taxonomic relations from texts. In Workshop on Ontology Learning and Population (associated to ECAI 2008), pages 1–6. University of Patras, juillet 2008.

[25] Franz Baader and Werner Nutt. Basic description logics. In The Description Logic Handbook, pages 43–95. Cambridge University Press, New York, NY, USA, 2003.

[26] Didier Bourigault, Nathalie Aussenac-Gilles, and Jean Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle, 18(1):87–110, 2004.

[27] Didier Bourigault and Nathalie Aussenac-Gilles. Construction d'ontologies à partir de textes. In Conférence Annuelle sur le Traitement Automatique des Langues, volume 2, pages 27–50, 2003.

[28] Mike Uschold and Michael Gruninger. Ontologies: principles, methods and applications. The Knowledge Engineering Review, 11:93–136, 1996.

[29] B. Chandrasekaran, John R. Josephson, and V. Richard Benjamins. What are ontologies, and why do we need them? IEEE Intelligent Systems, 14(1):20–26, January 1999.

[30] Olivier Bodenreider. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearbook of Medical Informatics, pages 67–79, 2008.

[31] Daniel L. Rubin, Nigam H. Shah, and Natalya F. Noy. Biomedical ontologies: a functional perspective. Briefings in Bioinformatics, 9(1):75–90, 2008.

[32] Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics, 16(6):1069–1080, 2015.

[33] Alan R Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3):229–236, 2010.

[34] Andon Tchechmedjiev, Amine Abdaoui, Vincent Emonet, Stella Zevio, and Clement Jonquet. SIFR annotator: ontology-based semantic annotation of french biomedical text and clinical notes. BMC Bioinformatics, 19(1):405, Nov 2018.

[35] Evelyn Camon, Michele Magrane, Daniel Barrell, Vivian Lee, Emily Dimmer, John Maslen, David Binns, Nicola Harte, Rodrigo Lopez, and Rolf Apweiler. The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Research, 32:D262–D266, January 2004.

[36] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, et al. Gene Ontology: tool for the unification of biology. Nature Genetics, 25:25, May 2000.

[37] Jagdev Bhogal, Andrew Macfarlane, and Peter WH. Smith. A review of ontology based query expansion. Information Processing & Management, 43(4):866–886, 2007.

[38] Chintan Patel, James Cimino, Julian Dolby, Achille Fokoue, Aditya Kalyanpur, Aaron Kershenbaum, Li Ma, Edith Schonberg, and Kavitha Srinivas. Matching patient records to clinical trials using ontologies. In Proceedings of the International Semantic Web Conference, pages 816–829, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[39] Paolo Besana, Marc Cuggia, Oussama Zekri, Annabel Bourde, and Anita Burgun. Using semantic web technologies for clinical trial recruitment. In Proceedings of the International Semantic Web Conference, pages 34–49, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[40] Emek Demir, Michael P Cary, Suzanne Paley, Ken Fukuda, Christian Lemer, Imre Vastrik, Guanming Wu, Peter D'Eustachio, Carl Schaefer, Joanne Luciano, et al. The BioPAX community standard for pathway data sharing. Nature Biotechnology, 28:935, September 2010.

[41] Jesùs Pardillo and Jose-Norberto Mazòn. Using ontologies for the design of data warehouses. International Journal of Database Management Systems, 3(2), 2011.

[42] Dominic Girardi, Johannes Dirnberger, and Michael Giretzlehner. An ontology-based clinical data warehouse for scientific research. Safety in Health, 1(1):6, Jul 2015.

[43] Robert Stevens, Patricia Baker, Sean Bechhofer, Gary Ng, Alex Jacoby, Norman W. Paton, Carole A. Goble, and Andy Brass. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics, 16(2):184–186, 02 2000.

[44] Antonio Sala and Sonia Bergamaschi. A mediator-based approach to ontology generation and querying of molecular and phenotypic cereals data. International Journal of Metadata, Semantics and Ontologies, 4(1/2):85–92, May 2009.

[45] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. PLoS Computational Biology, 5(7):1–12, 07 2009.

[46] Pietro H. Guzzi, Marco Mina, Concettina Guerra, and Mario Cannataro. Semantic similarity analysis of protein data: assessment with biological features and issues. Briefings in Bioinformatics, 13(5):569–585, 2012.

[47] Gaston K. Mazandu, Emile R. Chimusa, and Nicola J. Mulder. Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Briefings in Bioinformatics, 18(5):886–901, 2017.

[48] Stuart G. Jantzen, Ben JG Sutherland, David R. Minkley, and Ben F. Koop. GO trimming: systematically reducing redundancy in large gene ontology datasets. BMC Research Notes, 4(1):267, Jul 2011.

[49] Leila Ahmadian, Mariette van Engen-Verheul, Ferishta Bakhshi-Raiez, Niels Peek, Ronald Cornet, and Nicolette F. de Keizer. The role of standardized data and terminological systems in computerized clinical decision support systems: literature review and survey. International Journal of Medical Informatics, 80(2):81–93, February 2011.

[50] Adela Grando, Susan Farrish, Cynthia Boyd, and Aziz Boxwala. Ontological approach for safe and effective polypharmacy prescription. AMIA Annual Symposium Proceedings, 2012:291–300, November 2012.

[51] Tiffani J. Bright, E. Yoko Furuya, Gilad J. Kuperman, James J. Cimino, and Suzanne Bakken. Development and evaluation of an ontology for guiding appropriate antibiotic prescribing. Journal of Biomedical Informatics, 45(1):120 – 128, 2012.

[52] Daniele Nardi and Ronald J. Brachman. An introduction to description logics. In The Description Logic Handbook, pages 1–40. Cambridge University Press, New York, NY, USA, 2003.

[53] Katy Wolstencroft, Phillip Lord, Lydia Tabernero, Andrew Brass, and Robert Stevens. Protein classification using ontology classification. Bioinformatics, 22(14):e530–e538, 07 2006.

[54] Daniel L Rubin, Olivier Dameron, and Mark A Musen. Use of description logic classification to reason about consequences of penetrating injuries. AMIA Annual Symposium Proceedings, 2005:649–653, 2005.

[55] Matthew Horridge, Bijan Parsia, Natalya F Noy, and Mark A Musenm. Reasoning based quality assurance of medical ontologies: a case study. AMIA Annual Symposium Proceedings, 2014:671–680, November 2014.

[56] Rainer Winnenburg, Jonathan M. Mortensen, and Olivier Bodenreider. Using description logics to evaluate the consistency of drug-class membership relations in NDF-RT. Journal of Biomedical Semantics, 6(1):13, Mar 2015.

[57] Jean Noël Nikiema, Vianney Jouhet, and Fleur Mougin. Integrating cancer diagnosis terminologies based on logical definitions of SNOMED CT concepts. Journal of Biomedical Informatics, 74:46–58, October 2017.

[58] Mariano Fernández-López and Asunción Gómez-Pérez. Overview and analysis of methodologies for building ontologies. The Knowledge Engineering Review, 17(2):129–156, June 2002.

[59] Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. METHONTOLOGY: from ontological art towards ontological engineering. In Proceedings of the AAAI97 Spring Symposium, pages 33–40, Stanford, USA, March 1997.

[60] Natalya F. Noy and Deborah L. McGuinness. Ontology Development 101: A guide to creating your first ontology. Technical report, Stanford Medical Informatics, 2001.

[61] Bruno Bachimont, Antoine Isaac, and Raphaël Troncy. Semantic commitment for designing ontologies: a proposal. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, pages 114–121, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

[62] York Sure, Steffen Staab, and Rudi Studer. On-To-Knowledge Methodology (OTKM). Handbook on Ontologies, pages 117–132, 2004.

[63] Oscar Corcho, Mariano Fernández-López, and Asunción Gómez-Pérez. Ontological engineering: principles, methods, tools and languages. In Ontologies for Software Engineering and Software Technology, pages 1–48. Springer Berlin Heidelberg, 2006.

[64] Bhavna Orgun, Mark Dras, Abhaya Nayak, and Geoff James. Approaches for semantic interoperability between domain ontologies. Expert Systems, 25(3):179–196, 7 2008.

[65] Fleur Mougin. Conception d'un modèle Web sémantique appliqué à la génomique fonctionnelle. Thèse, Université de Rennes 1, 2006.

[66] Fleur Mougin, Anita Burgun, Olivier Bodenreider, Julie Chabalier, Olivier Loréal, and Pierre Le Beux. Automatic methods for integrating biomedical data sources in a mediator-based system. In Data Integration in the Life Sciences, pages 61–76, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[67] Fleur Mougin, Anita Burgun-Parenthoine, Olivier Loréal, and Pierre Le Beux. Towards the automatic generation of biomedical sources schema. Studies in Health Technology and Informatics, 107:783–787, 2004.

[68] Fleur Mougin, Anita Burgun-Parenthoine, and Olivier Bodenreider. Mapping data elements to terminological resources for integrating biomedical data sources. BMC Bioinformatics, 7:S6, 2006.

[69] Fleur Mougin, Anita Burgun-Parenthoine, and Olivier Bodenreider. Using WordNet to improve the mapping of data elements to UMLS for data sources integration. AMIA Annual Symposium Proceedings, pages 574–578, 2006.

[70] Fleur Mougin, Julie Chabalier, Olivier Bodenreider, and Anita Burgun-Parenthoine. Méthodes de mapping situées aux niveaux instance et schéma pour l'intégration de sources de données hétérogènes. In Actes des journées francophones d'Ingénierie des Connaissances (IC), pages 37–49, 2007.

[71] Philippe Laublet, Chantal Reynaud, and Jean Charlet. Sur quelques aspects du web sémantique. In Actes des deuxièmes assises du GdR I3, pages 59–78. Cépaduès Editions, 2002.

[72] Hacid Mohand-Said and Reynaud Chantal. L'intégration de sources de données. Revue I3 - Information Interaction Intelligence, 2004.

[73] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientific American, 284(5):34–43, May 2001.

[74] Paul Avillach, Fleur Mougin, Michel Joubert, Frantz Thiessard, Antoine Pariente, Jean-Charles Dufour, Gianluca Trifirò, Giovanni Polimeni, Maria Antonietta Catania, Carlo Giaquinto, and EU-ADR consortium. A semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project. Studies in Health Technology and Informatics, 150:190–194, 2009.

[75] Paul Avillach, Fleur Mougin, Michel Joubert, Frantz Thiessard, Antoine Pariente, Jean-Charles Dufour, and Marius Fieschi. Approche sémantique pour l'identification homogène d'effets indésirables et de médicaments dans huit bases de données de patients : une contribution au projet européen ALERT. Informatique et Santé, 17:251–261, 2009.

[76] Paul Avillach, Michel Joubert, Frantz Thiessard, Gianluca Trifirò, Jean-Charles Dufour, Antoine Pariente, Fleur Mougin, Giovanni Polimeni, Maria Antonietta Catania, Carlo Giaquinto, and EU-ADR consortium. Design and evaluation of a semantic approach for the homogeneous identification of events in eight patient databases: a contribution to the European EU-ADR project. Studies in Health Technology and Informatics, 160(Pt 2):1085–1089, 2010.

[77] Paul Avillach, Preciosa M Coloma, Rosa Gini, Martijn Schuemie, Fleur Mougin, Jean-Charles Dufour, Giampiero Mazzaglia, Carlo Giaquinto, Carla Fornari, Ron Herings, and EU-ADR consortium. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. Journal of the American Medical Informatics Association, 20(1):184–192, January 2013.

[78] Fleur Mougin and Olivier Bodenreider. Approaches to eliminating cycles in the UMLS Metathesaurus: naïve vs. formal. AMIA Annual Symposium proceedings, pages 550–554, 2005.

[79] Anita Burgun, Fleur Mougin, and Olivier Bodenreider. Two approaches to integrating phenotype and clinical information. AMIA Annual Symposium Proceedings, 2009:75–79, 2009.

[80] Fleur Mougin, Anita Burgun, and Olivier Bodenreider. Comparing drug-class membership in ATC and NDF-RT. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, IHI'12, pages 437–444, New York, NY, USA, 2012. ACM.

[81] Fleur Mougin and Olivier Bodenreider. Auditing the NCI thesaurus with semantic web technologies. AMIA Annual Symposium Proceedings, pages 500–504, November 2008.

[82] Fleur Mougin, Olivier Bodenreider, and Anita Burgun. Analyzing polysemous concepts from a clinical perspective: application to AUDITING concept categorization in the UMLS. Journal of Biomedical Informatics, 42(3):440–451, June 2009.

[83] Khadim Dramé. Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical. Thèse, Université de Bordeaux, December 2014.

[84] Khadim Dramé, Gayo Diallo, and Fleur Mougin. Towards a bilingual Alzheimer's disease terminology acquisition using a parallel corpus. Proceedings of the 24th Conference of Medical Informatics in Europe, 180:179–183, 2012.

[85] Khadim Dramé, Gayo Diallo, and Fleur Mougin. Construction d'une ontologie bilingue de la maladie d'Alzheimer à partir de textes médicaux. In Atelier IC pour l'Interopérabilité Sémantique dans les applications en e-Santé, 2012.

[86] Khadim Dramé, Gayo Diallo, Fleur Delva, Jean François Dartigues, Evelyne Mouillet, Roger Salamon, and Fleur Mougin. Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology: an application to Alzheimer's disease. Journal of Biomedical Informatics, 48:171–182, April 2014.

[87] Khadim Dramé, Fleur Mougin, and Gayo Diallo. Query expansion using external resources for improving information retrieval in the biomedical domain. In CLEF (Working Notes), pages 189–194, 2014.

[88] Khadim Dramé, Fleur Mougin, and Gayo Diallo. A k-nearest neighbor based method for improving large scale biomedical document indexing. In Proceedings of the 6th International Symposium on Semantic Mining in Biomedicine (SMBM), October 2014.

[89] Khadim Dramé, Fleur Mougin, and Gayo Diallo. Large scale biomedical texts classification: a kNN and an ESA-based approaches. Journal of Biomedical Semantics, 7:40, June 2016.

[90] Fleur Mougin, David Auber, Romain Bourqui, Gayo Diallo, Isabelle Dutour, Vianney Jouhet, Frantz Thiessard, Rodolphe Thiébaut, and Patricia Thébault. Visualizing omics and clinical data: Which challenges for dealing with their variety? Methods, 132:3–18, January 2018.

[91] Aarón Ayllón-Benítez, Fleur Mougin, Julien Allali, Rodolphe Thiébaut, and Patricia Thébault. A new method for evaluating the impacts of semantic similarity measures on the annotation of gene sets. PLoS One, 11:e0208037, November 2018.

[92] Aarón Ayllón-Benítez, Romain Bourqui, Patricia Thébault, and Fleur Mougin. GSAn: an alternative to enrichment analysis for annotating gene sets. Nucleic Acids Research, 2019. Submitted.

[93] Aarón Ayllón-Benítez, Fleur Mougin, Jesualdo Tomás Fernández-Breis, Manuel Quesada Martinez, Patricia Thébault, and Romain Bourqui. Deciphering gene sets annotations with ontology based visualization. In Proceedings of 21st International Conference on Information Visualization, pages 170–175, London, United Kingdom, July 2017.

[94] Vianney Jouhet. Adaptation automatique des données de prises en charge hospitalières pour une utilisation secondaire en cancérologie. Thèse, Université de Bordeaux, December 2016.

[95] Vianney Jouhet, Fleur Mougin, Bérénice Brechat, and Frantz Thiessard. Building a model for disease classification integration in oncology, an approach based on the National Cancer Institute thesaurus. Journal of Biomedical Semantics, 8(1):6, February 2017.

[96] Jean Noël Nikiema, Fleur Mougin, and Vianney Jouhet. Finding the appropriate transversal relations between entities of distinct knowledge resources: a step forward on semantic integration. Journal of Biomedical Semantics, 2019. Submitted.

[97] Jean Noël Nikiema, Fleur Mougin, and Vianney Jouhet. Processus de prétraitement des libellés d'une terminologie d'interface. In Actes du symposium sur l'Ingénierie de l'Information Médicale, pages 95–103, November 2017.

[98] Fleur Mougin, Marie Dupuch, and Natalia Grabar. Improving the mapping between MedDRA and SNOMED CT. In Artificial Intelligence in Medicine, number 6747 in Lecture Notes in Computer Science, pages 220–224. Springer Berlin Heidelberg, January 2011.

[99] Fleur Mougin and Natalia Grabar. Using a cross-language approach to acquire new mappings between two biomedical terminologies. In Artificial Intelligence in Medicine, number 7885 in Lecture Notes in Computer Science, pages 221–226. Springer Berlin Heidelberg, January 2013.

[100] Frantz Thiessard, Fleur Mougin, Gayo Diallo, Vianney Jouhet, Sébastien Cossin, Nicolas Garcelon, Boris Campillo, Wassim Jouini, Julien Grosjean, Philippe Massari, et al. RAVEL: retrieval and visualization in electronic health records. Studies in Health Technology and Informatics, 180:194–198, 2012.

[101] Natalia Grabar, Marie Dupuch, and Fleur Mougin. Dommages collatéraux de la fusion de terminologies. In Proceedings of the 9th International Conference on Terminology and Artificial Intelligence, pages 10–16, 2011.

[102] Fleur Mougin and Natalia Grabar. Auditing the multiply-related concepts within the UMLS. Journal of the American Medical Informatics Association, 21(e2):e185–193, October 2014.

[103] Thierry Hamon, Natalia Grabar, and Fleur Mougin. Natural language question analysis for querying biomedical Linked Data. In Proceedings of the ISWC Workshop Natural Language Interfaces for Web of Data (NLIWoD), volume 8, October 2014.

[104] Thierry Hamon, Fleur Mougin, and Natalia Grabar. Generating and Executing Complex Natural Language Queries across Linked Data. Studies in Health Technology and Informatics, 216:815–820, 2015.

[105] Thierry Hamon, Natalia Grabar, and Fleur Mougin. Querying biomedical Linked Data with natural language questions. Semantic Web, 8(4):581–599, January 2017.

[106] Thierry Hamon, Natalia Grabar, Fleur Mougin, and Frantz Thiessard. Description of the POMELO system for the task 2 of QALD-2014. In CLEF (Working Notes), volume 1180 of CEUR Workshop Proceedings, pages 1212–1223. CEUR-WS.org, 2014.

[107] Thierry Hamon, Vincent Tabanou, Fleur Mougin, Natalia Grabar, and Frantz Thiessard. POMELO: MEDLINE corpus with manually annotated food-drug interactions. In Proceedings of the Biomedical NLP Workshop associated with Recent Advances in Natural Language Processing, pages 73–80, September 2017.

[108] Georgeta Bordea, Frantz Thiessard, Thierry Hamon, and Fleur Mougin. Automatic query selection for acquisition and discovery of food-drug interactions. In Experimental IR Meets Multilinguality, Multimodality, and Interaction, pages 115–120, Cham, 2018. Springer International Publishing.

[109] Janez Brank, Marko Grobelnik, and Dunja Mladenić. A survey of ontology evaluation techniques. In Proceeding of 8th International multi-conference of Information Society, pages 166–169, 2005.

[110] Stefan Schlobach and Ronald Cornet. Non-standard reasoning services for the debugging of description logic terminologies. In Proceedings of the 18th International Joint Conference on Artificial Intelligence, IJCAI'03, pages 355–360, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc.

[111] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou. Ontology change: classification and survey. The Knowledge Engineering Review, 23(2):117–152, June 2008.

[112] Asunción Gómez-Pérez. Evaluation of ontologies. International Journal of Intelligent Systems, 16(3):391–409, 2001.

[113] Christopher Welty and Nicola Guarino. Supporting ontological analysis of taxonomic relationships. Data & Knowledge Engineering, 39(1):51–74, October 2001.

[114] Xinxin Zhu, Jung-Wei Fan, David M. Baorto, Chunhua Weng, and James J. Cimino. A review of auditing methods applied to the content of controlled biomedical terminologies. Journal of Biomedical Informatics, 42(3):413–425, June 2009.

[115] Asunción Gómez-Pérez. Towards a framework to verify knowledge sharing technology. Expert Systems with Applications, 11(4):519 – 529, 1996.

[116] Denny Vrandečić. Ontology evaluation. Handbook on Ontologies, pages 293–313, 2009.

[117] Muhammad Amith, Zhe He, Jiang Bian, Juan Antonio Lossio-Ventura, and Cui Tao. Assessing the practice of biomedical ontology evaluation: gaps and opportunities. Journal of Biomedical Informatics, 80:1–13, 2018.

[118] Pierre Zweigenbaum. L'UMLS entre langue et ontologie: une approche pragmatique dans le domaine médical. Revue d'Intelligence Artificielle, 18(1):111–137, 2004.

[119] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32 Database issue:D267–70, 2004.

[120] Natalya Fridman Noy and Mark A. Musen. An algorithm for merging and aligning ontologies: Automation and tool support. In Proceedings of the Workshop on Ontology Management at the Sixteenth National Conference on Artificial Intelligence, 1999.

[121] Alexa T. McCray. An upper-level ontology for the biomedical domain. Comparative and Functional Genomics, 4(1):80–84, 2003.

[122] Olivier Bodenreider and Alexa T. McCray. Exploring semantic groups through visual approaches. Journal of Biomedical Informatics, 36(6):414 – 432, 2003. Unified Medical Language System.

[123] James Pustejovsky. The generative lexicon. Computational Linguistics, 17(4):409–441, December 1991.

[124] Alexa T. McCray, Anita Burgun, and Olivier Bodenreider. Aggregating UMLS semantic types for reducing conceptual complexity. Studies in Health Technology and Informatics, 84(pt 1):216–220, 2001.

[125] Pierre Zweigenbaum, Bruno Bachimont, Jacques Bouaud, Jean Charlet, and Jean-François Boisvieux. Issues in the structuring and acquisition of an ontology for medical language understanding. Methods of Information in Medicine, 34(1-2):15–24, March 1995.

[126] James J. Cimino, Hua Min, and Yehoshua Perl. Consistency across the hierarchies of the UMLS semantic network and metathesaurus. Journal of Biomedical Informatics, 36(6):450–461, December 2003.

[127] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, and Lawrence W. Wright. NCI thesaurus: A semantic model integrating cancer-related clinical and molecular information. Journal of Biomedical Informatics, 40(1):30–43, February 2007.

[128] Amy Y. Wang, James W. Barrett, Tim Bentley, David Markwell, Colin Price, Kent A. Spackman, and Michael Q. Stearns. Mapping between SNOMED RT and Clinical Terms version 3: a key component of the SNOMED CT development process. AMIA Annual Symposium Proceedings, pages 741–745, 2001.

[129] Kin Wah Fung, William T. Hole, Stuart J. Nelson, Suresh Srinivasan, Tammy Powell, and Laura Roth. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. Journal of the American Medical Informatics Association, 12(4):486–494, August 2005.

[130] Gary H. Merrill. The MedDRA paradox. AMIA Annual Symposium Proceedings, pages 470–474, 2008.

[131] Olivier Bodenreider, Anita Burgun, and Thomas C. Rindflesch. Assessing the consistency of a biomedical terminology through lexical knowledge. International Journal of Medical Informatics, 67(1):85 – 95, 2002.

[132] Huanying Gu, Yehoshua Perl, Gai Elhanan, Hua Min, Li Zhang, and Yi Peng. Auditing concept categorizations in the UMLS. Artificial Intelligence in Medicine, 31(1):29 – 44, 2004.

[133] Huanying Gu, Gai Elhanan, Michael Halper, and Zhe He. Questionable relationship triples in the UMLS. In Proceedings of 2012 IEEE International Conference on Biomedical and Health Informatics, pages 713–716, January 2012.

[134] Frank W. Hartel, Sherri de Coronado, Robert Dionne, Gilberto Fragoso, and Jennifer Golbeck. Modeling a description logic vocabulary for cancer research. Journal of Biomedical Informatics, 38(2):114–129, April 2005.

[135] Fleur Mougin. Identifying redundant and missing relations in the Gene Ontology. Studies in Health Technology and Informatics, 210:195–199, 2015.

[136] Philip V. Ogren, K. Bretonnel Cohen, George K. Acquaah-Mensah, Jens Eberlein, and Lawrence Hunter. The compositional structure of Gene Ontology terms. In Proceedings of the Pacific Symposium on Biocomputing, pages 214–225, 2004.

[137] Natalia Grabar, Cédric Bousquet, and Marie-Christine Jaulent. Le traitement automatique des langues et la fouille des textes en biologie, un nouveau défi pour Gene Ontology. Revue I3 - Information Interaction Intelligence, 7(1), 2006.

[138] Guo-Qiang Zhang and Olivier Bodenreider. Using SPARQL to test for lattices: application to quality assurance in biomedical ontologies. Proceedings of the International Conference on Semantic Web, 6497:273–288, 2010.

[139] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. The incredible ELK: from polynomial procedures to efficient reasoning with $\mathcal{EL}$ ontologies. Journal of Automated Reasoning, 53:1–61, 2013.

[140] Jesualdo Tomás Fernández-Breis, Luigi Iannone, Ignazio Palmisano, Alan L. Rector, and Robert Stevens. Enriching the Gene Ontology via the dissection of labels using the ontology pre-processor language. In Knowledge Engineering and Management by the Masses, pages 59–73, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[141] Christopher J. Mungall, Michael Bada, Tanya Z. Berardini, Jennifer Deegan, Amelia Ireland, Midori A. Harris, David P. Hill, and Jane Lomax. Cross-product extensions of the Gene Ontology. Journal of Biomedical Informatics, 44(1):80 – 86, 2011.

[142] David P. Hill, Nico Adams, Mike Bada, Colin Batchelor, Tanya Z. Berardini, Heiko Dietze, Harold J. Drabkin, Marcus Ennis, Rebecca E. Foulger, Midori A. Harris, et al. Dovetailing biology and chemistry: integrating the Gene Ontology with the ChEBI chemical ontology. BMC Genomics, 14(1):513, Jul 2013.

[143] Manuel Quesada-Martínez, Jesualdo Tomás Fernández-Breis, Robert Stevens, and Nathalie Aussenac-Gilles. OntoEnrich: a platform for the lexical analysis of ontologies. In Knowledge Engineering and Knowledge Management, pages 172–176, Cham, 2015. Springer International Publishing.

[144] Manuel Quesada-Martínez, Eleni Mikroyannidi, Jesualdo Tomás Fernández-Breis, and Robert Stevens. Approaching the axiomatic enrichment of the gene ontology from a lexical perspective. Artificial Intelligence in Medicine, 65(1):35 – 48, 2015.

[145] Rashmie Abeysinghe, Xufeng Qu, and Cui Licong. Identifying similar non-lattice subgraphs in Gene Ontology based on structural isomorphism and semantic similarity of concept labels. AMIA Annual Symposium proceedings, 2018:1186–1195, December 2018.

[146] Licong Cui, Wei Zhu, Shiqiang Tao, Guo-Qiang Zhang, James T Case, and Olivier Bodenreider. Mining non-lattice subgraphs for detecting missing hierarchical relations and concepts in SNOMED CT. Journal of the American Medical Informatics Association, 24(4):788–798, 02 2017.

[147] Rashmie Abeysinghe, Michael A Brooks, Jeffery Talbert, and Cui Licong. Quality assurance of NCI thesaurus by mining structural-lexical patterns. AMIA Annual Symposium proceedings, 2017:364–373, April 2018.

[148] Guoqian Jiang and Christopher G Chute. Auditing the semantic completeness of SNOMED CT using formal concept analysis. Journal of the American Medical Informatics Association, 16(1):89–102, 2009.

[149] Guo-Qiang Zhang and Olivier Bodenreider. Large-scale, exhaustive lattice-based structural auditing of SNOMED CT. AMIA Annual Symposium Proceedings, 2010:922–926, November 2010.

[150] Peter Haase and Guilin Qi. An analysis of approaches to resolving inconsistencies in DL-based ontologies. In Proceedings of the International Workshop on Ontology Dynamics (IWOD), pages 97–109, 2007.

[151] Qiu Ji, Zhiqiang Gao, Zhisheng Huang, and Man Zhu. Measuring effectiveness of ontology debugging systems. Knowledge-Based Systems, 71:169 – 186, 2014.

[152] Jonathan M. Mortensen. Crowdsourcing ontology verification. In Proceedings of the International Conference on Semantic Web Conference, pages 448–455, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[153] Jonathan M Mortensen, Evan P Minty, Michael Januszyk, Timothy E Sweeney, Alan L Rector, Natalya F Noy, and Mark A Musen. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. Journal of the American Medical Informatics Association, 22(3):640–648, May 2015.

[154] Hua Min, Yehoshua Perl, Yan Chen, Michael Halper, James Geller, and Yue Wang. Auditing as part of the terminology design life cycle. Journal of the American Medical Informatics Association, 13(6):676–690, 11 2006.

[155] Natalya Fridman Noy and Carole D. Hafner. The state of the art in ontology design: A survey and comparative review. AI Magazine, 18:53–74, 1997.

[156] Astrid Duque-Ramos, Jesualdo Tomás Fernández-Breis, Robert Stevens, and Nathalie Aussenac-Gilles. OQuaRE: a SQuaRE-based approach for evaluating the quality of ontologies. Journal of Research and Practice in Information Technology, 43(2):159–176, 2011.

[157] Jérôme Euzenat, Christian Meilicke, Heiner Stuckenschmidt, Pavel Shvaiko, and Cássia Trojahn. Ontology Alignment Evaluation Initiative: Six Years of Experience, pages 158–192. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[158] Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: The state of the art. The Knowledge Engineering Review, 18(1):1–31, January 2003.

[159] Ying Ding, Dieter Fensel, Michel Klein, and Borys Omelayenko. The semantic web: yet another hip? Data & Knowledge Engineering, 41(2-3):205–227, June 2002.

[160] H. Sofia Pinto, Asunción Gómez-Pérez, and João P. Martins. Some issues on ontology integration. In Proc. of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends, 18, 1999.

[161] Zharko Aleksovski, Michel Klein, Warner ten Kate, and Frank van Harmelen. Exploiting the structure of background knowledge used in ontology matching. In Proceedings of the ISWC Ontology Matching Workshop, pages 13–24, 2006.

[162] Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, Ian Horrocks, and Rafael Berlanga. Ontology integration using mappings: towards getting the right logical consequences. In The Semantic Web: Research and Applications, pages 173–187, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[163] Jérôme Euzenat and Pavel Shvaiko. Ontology Matching. Springer-Verlag, Berlin, Heidelberg, 2007.

[164] Pavel Shvaiko and Jérôme Euzenat. Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering, 25(1):158–176, January 2013.

[165] Jérôme Euzenat and Pavel Shvaiko. Ontology Matching, Second Edition. Springer, 2013.

[166] Lorena Otero-Cerdeira, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez. Ontology matching: a literature review. Expert Systems with Applications, 42(2):949–971, 2015.

[167] Daniel Faria, Catia Pesquita, Isabela Mott, Catarina Martins, Francisco M. Couto, and Isabel F. Cruz. Tackling the challenges of matching biomedical ontologies. Journal of Biomedical Semantics, 9(1):4, Jan 2018.

[168] Matthew Horridge, Bijan Parsia, and Ulrike Sattler. The state of bio-medical ontologies. In Proceedings of Bio-Ontologies 2011, 2011.

[169] Anika Oellrich, Dietrich Rebholz-Schuhmann, Georgios V. Gkoutos, Michel Dumontier, Paul Schofield, Robert Hoehndorf, and Sarala Wimalaratne. A common layer of interoperability for biomedical ontologies based on OWL EL. Bioinformatics, 27(7):1001–1008, 02 2011.

[170] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. The VLDB Journal, 10(4):334–350, Dec 2001.

[171] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. Journal on Data Semantics, 4:146–171, 2005.

[172] David W Bates, R Scott Evans, Harvey Murff, Peter D Stetson, Lisa Pizziferri, and George Hripcsak. Detecting adverse events using information technology. Journal of the American Medical Informatics Association, 10(2):115–128, 2003.

[173] Gianluca Trifirò, Antoine Pariente, Preciosa M. Coloma, Jan A. Kors, Giovanni Polimeni, Ghada Miremont-Salamé, Maria Antonietta Catania, Francesco Salvo, Anaelle David, Nicholas Moore, and EU-ADR consortium. Data mining on electronic health record databases for signal detection in pharmacovigilance: which events to monitor? Pharmacoepidemiology and Drug Safety, 18(12):1176–1184, December 2009.

[174] Elliot G. Brown, Louise Wood, and Sue Wood. The Medical Dictionary for Regulatory Activities (MedDRA). Drug Safety, 20(2):109–117, Feb 1999.

[175] Michael Q Stearns, Colin Price, Kent A Spackman, and Amy Y Wang. SNOMED clinical terms: overview of the development process and project status. AMIA Annual Symposium Proceedings, pages 662–666, 2001.

[176] David J Rothwell, Roger A Cote, Jean-François Cordeau, and Maryse Boisvert. Developing a standard data structure for medical language–the SNOMED proposal. Proceedings of the Symposium on Computer Applications in Medical Care, pages 695–699, 1993.

[177] Robert H Dolin, Kent A Spackman, and David Markwell. Selective retrieval of pre- and post-coordinated SNOMED concepts. AMIA Annual Symposium Proceedings, pages 210–214, 2002.

[178] Natalia Grabar and Pierre Zweigenbaum. A general method for sifting linguistic knowledge from structured terminologies. AMIA Annual Symposium Proceedings, pages 310–314, 2000.

[179] Natalia Grabar, Paul-Christophe Varoutas, Philippe Rizand, Alain Livartowski, and Thierry Hamon. Automatic acquisition of synonym resources and assessment of their impact on the enhanced search in EHRs. Methods of Information in Medicine, 48:149–54, 02 2009.

[180] William T. Hole and Suresh Srinivasan. Discovering missed synonymy in a large concept-oriented metathesaurus. AMIA Annual Symposium Proceedings, pages 354–358, 2000.

[181] Calum Muir and Constance Percy. Classification and coding of neoplasms. Cancer registration: principles and methods, 95:64–81, 1991.

[182] Bérénice Brechat, Fleur Mougin, Frantz Thiessard, and Vianney Jouhet. Mapping de terminologies diagnostiques en cancérologie par l'intermédiaire du NCI metathesaurus. In Actes des 15èmes Journées Francophones d'Informatique Médicale, pages 34–43, 2014.

[183] Cui Tao, Jyotishman Pathak, Harold R. Solbrig, Wei-Qi Wei, and Christopher G. Chute. Terminology representation guidelines for biomedical ontologies in the semantic web notations. Journal of Biomedical Informatics, 46(1):128–138, February 2013.

[184] Jean Noël Nikiema, Fleur Mougin, and Vianney Jouhet. Utilisation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie. In Actes des 6èmes Journées Francophones sur les Ontologies, pages 142–148, 2016.

[185] Jean Noël Nikiema, Vianney Jouhet, and Fleur Mougin. Evaluation de la SNOMED CT comme support à l'alignement de terminologies diagnostiques en cancérologie. In Atelier IC pour IA & Santé aux 27èmes journées francophones d'Ingénierie des Connaissances, 2016.

[186] Brigitte Safar, Chantal Reynaud, and François Calvier. Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire. In Actes des 1ères journées francophones sur les ontologies, pages 21–35, 2007.

[187] Yevgeny Kazakov, Markus Krötzsch, and František Simančík. ELK reasoner: architecture and evaluation. In Proceedings of the OWL Reasoner Evaluation Workshop 2012. CEUR Workshop Proceedings, July 2012.

[188] Nicola Guarino. Some ontological principles for designing upper level lexical resources. In First International Conference on Language Resources and Evaluation, pages 527–534, 1998.

[189] Anand Kumar and Barry Smith. Oncology ontology in the NCI thesaurus. In Artificial Intelligence in Medicine, pages 213–220, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[190] Patrick Lambrix and He Tan. SAMBO - A System for Aligning and Merging Biomedical Ontologies. Web Semantics: Science, Services and Agents on the World Wide Web, 4(3), 2006.

[191] Elena Simperl. Reusing ontologies on the semantic web: a feasibility study. Data & Knowledge Engineering, 68(10):905 – 925, 2009.

[192] Christoph Quix, Pratanu Roy, and David Kensche. Automatic selection of background knowledge for ontology matching. In Proceedings of the International Workshop on Semantic Web Information Management, pages 5:1–5:7, New York, NY, USA, 2011. ACM.

[193] Daniel Faria, Catia Pesquita, Emanuel Santos, Isabel F Cruz, and Francisco M Couto. Automatic background knowledge selection for matching biomedical ontologies. PLoS One, 9(11):e111226–e111226, November 2014.

[194] Asuncion Gómez-Pérez and David Manzano-Macho. A survey of ontology learning methods and techniques. Deliverable 1.5, OntoWeb Consortium, 2003.

[195] Lina Zhou. Ontology learning: state of the art and open issues. Information Technology and Management, 8(3):241–252, Sep 2007.

[196] Amal Zouaq and Roger Nkambou. A Survey of Domain Ontology Engineering: Methods and Tools, pages 103–119. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[197] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: an overview. In Ontology Learning from Text: Methods, Evaluation and Applications, volume 123, pages 3–12. Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

[198] Wilson Wong, Wei Liu, and Mohammed Bennamoun. Ontology learning from text: a look back and into the future. ACM Computing Surveys, 44(4):20:1–20:36, September 2012.

[199] Nicola Guarino, Claudio Masolo, and Guido Vetere. OntoSeek: content-based access to the web. IEEE Intelligent Systems and their Applications, 14(3):70–80, May 1999.

[200] Didier Bourigault and Cécile Fabre. Approche linguistique pour l'analyse syntaxique de corpus. Cahiers de Grammaire, 25:131–151, 2000.

[201] Claude Chrisment, Ollivier Haemmerlé, Nathalie Hernandez, and Josiane Mothe. Méthodologie de transformation d'un thesaurus en une ontologie de domaine. Revue d'Intelligence Artificielle, 22(1):7–37, 2008.

[202] I. Dan Melamed. A portable algorithm for mapping bitext correspondence. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, pages 305–312, Stroudsburg, PA, USA, 1997. Association for Computational Linguistics.

[203] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pages 177–180. Association for Computational Linguistics, 2007.

[204] Simon Jupp, Sean Bechhofer, and Robert Stevens. SKOS with OWL: don't be full-ish! In Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, 2008.

[205] Csongor Nyulas, Mark A. Musen, Natalya F. Noy, Nigam H. Shah, Paul R. Alexander, Tania Tudorache, and Patricia L. Whetzel. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39(suppl_2):W541–W545, 06 2011.

[206] Nathalie Aussenac-Gilles, Sylvie Despres, and Sylvie Szulman. The Terminae method and platform for ontology engineering from texts. In Ontology Learning and Population: Bridging the Gap between Text and Knowledge, volume 167, pages 192–223. Frontiers in Artificial Intelligence and Applications. IOS press, February 2008.

[207] Pierre Grenon and Barry Smith. SNAP and SPAN: Towards dynamic spatial ontology. Spatial Cognition & Computation, 4(1):69–104, 2004.

[208] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J Mungall, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 25(11):1251–5, November 2007.

[209] Janna Hastings, Werner Ceusters, Mark Jensen, Kevin Mulligan, and Barry Smith. Representing mental functioning : ontologies for mental health and disease. In Proceedings of the 3rd International Conference on Biomedical Ontology, 2012.

[210] Sumi Yoshikawa, Kenji Satou, and Akihiko Konagaya. Drug Interaction Ontology (DIO) for inferences of possible drug-drug interactions. Studies in Health Technology and Informatics, 107(Pt 1):454–458, 2004.

[211] María Herrero-Zazo, Janna Hastings, Isabel Segura-Bedmar, Samuel Croset, Paloma Martínez, and Christoph Steinbeck. An ontology for drug-drug interactions. In Semantic Web Applications and Tools for Healthcare and Life Sciences (SWAT4LS), volume 1114. CEUR Workshop Proceedings, 2013.

[212] Mathias Brochhausen, Jodi Schneider, Daniel Malone, Philip E. Empey, William R. Hogan, and Richard D. Boyce. Towards a foundational representation of potential drug-drug interaction knowledge. In 1st International Drug-Drug Interaction Knowledge Representation Workshop (DIKR 2014), volume 1309, pages 16–31. CEUR-WS, 2014.

[213] Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck.

ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research, 44(D1):D1214–D1219, January 2016.

[214] Damion M. Dooley, Emma J. Griffiths, Gurinder S. Gosal, Pier L. Buttigieg, Robert Hoehndorf, Matthew C. Lange, Lynn M. Schriml, Fiona S. L. Brinkman, and William W. L. Hsiao. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. Nature Partner Journals Science of Food, 2(1):23, December 2018.

[215] Junguk Hur, Arzucan Özgür, Zuoshuang Xiang, and Yongqun He. Development and application of an interaction network ontology for literature mining of vaccine-associated gene-gene interactions. Journal of Biomedical Semantics, 6:2–2, jan 2015.

[216] Mark A. Musen. The Protégé project: A look back and a look forward. AI Matters, 1(4):4–12, June 2015.

[217] GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990-2017: a systematic analysis for the global burden of disease study 2017. The Lancet, 2019.

[218] Axel Reymonet, Jérôme Thomas, and Nathalie Aussenac-Gilles. Modélisation de ressources termino-ontologiques en OWL. In Journées Francophones d'Ingénierie des Connaissances (IC 2007), pages 169–180, Grenoble, France, July 2007. Cépaduès Editions.

[219] Pierre-Yves Vandenbussche and Jean Charlet. Méta-modèle général de description de ressources terminologiques et ontologiques. In Journées Francophones d'Ingénierie des Connaissances (IC 2009), volume 20, Hammamet, Tunisia, May 2009.

[220] Philipp Cimiano, Paul Buitelaar, John McCrae, and Michael Sintek. LexInfo: A declarative model for the lexicon-ontology interface. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, 9(1), 2011.

[221] Spiros C. Denaxas. Integrating bio-ontologies and controlled clinical terminologies: from base pairs to bedside phenotypes. Methods in Molecular Biology, pages 275–287, 2017.

[222] Thomas Bleazard, Janine A Lamb, and Sam Griffiths-Jones. Bias in microRNA functional enrichment analysis. Bioinformatics, 31(10):1592–1598, 2015.

[223] Winston A. Haynes, Aurelie Tomczak, and Purvesh Khatri. Gene annotation bias impedes biomedical research. Scientific Reports, 8(1):1362, January 2018.

[224] Daniel Faria, Andreas Schlicker, Catia Pesquita, Hugo Bastos, Antonio EN. Ferreira, Mario Albrecht, and André O. Falcão. Mining GO Annotations for Improving Annotation Consistency. PLoS One, 7:e40519, July 2012.

[225] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The Reactome pathway knowledgebase. Nucleic Acids Research, 46(D1):D649–D655, 11 2017.

[226] Cesar Arze, Gang Feng, Mark Mazaitis, Suvarna Nadendla, Victor Felix, Yu-Wei Wayne Chang, Lynn Marie Schriml, and Warren Alden Kibbe. Disease Ontology: a backbone for disease semantic integration. Nucleic Acids Research, 40(D1):D940–D946, 11 2011.

[227] Xiaoli Jiao, Brad T. Sherman, Da Wei Huang, Robert Stephens, Michael W. Baseler, H Clifford Lane, and Richard A. Lempicki. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinformatics, 28(13):1805–1806, Jul 2012.

[228] François Belleau, Marc-Alexandre Nolin, Nicole Tourigny, Philippe Rigault, and Jean Morissette. Bio2RDF: Towards a mashup to build bioinformatics knowledge systems. Journal of Biomedical Informatics, 41(5):706–716, 2008.

[229] Matthias Samwald, Anja Jentzsch, Christopher Bouton, Claus Stie Kallesøe, Egon Willighagen, Janos Hajagos, M Scott Marshall, Eric Prud'hommeaux, Oktie Hassenzadeh, Elgar Pichler, and Susie Stephens. Linked open drug data for pharmaceutical research and development. Journal of Cheminformatics, 3(1):19, 2011.

[230] María Jesús García Godoy, Esteban López-Camacho, Ismael Navas-Delgado, and José F. Aldana-Montes. Sharing and executing linked data queries in a collaborative environment. Bioinformatics, 29(13):1663–1670, 2013.

[231] Alison Callahan, José Cruz-Toledo, and Michel Dumontier. Ontology-based querying with Bio2RDF's linked open data. Journal of Biomedical Semantics, 4 Suppl 1:S1, 4 2013.

[232] Ben Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In Proceedings of the IEEE Symposium on Visual languages, pages 336–343, 1996.

[233] Vincent Canuel, Bastien Rance, Paul Avillach, Philippe Degoulet, and Anita Burgun. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Briefings in Bioinformatics, 16(2):280–290, mar 2015.

[234] Sascha Herzinger, Wei Gu, Venkata Satagopam, Serge Eifes, Kavita Rege, Adriano Barbosa-Silva, Reinhard Schneider, and eTRIKS consortium. SmartR: an open-source platform for interactive visual analytics for translational research data. Bioinformatics, 33(14):2229–2231, 2017.

[235] Marek Dudáš, Steffen Lohmann, Vojtěch Svátek, and Dmitry Pavlov. Ontology visualization methods and tools: a survey of the state of the art. The Knowledge Engineering Review, 33:e10, 2018.

[236] Sasa Kuhar and Vili Podgorelec. Ontology visualization for domain experts: a new solution. In 16th International Conference on Information Visualisation, pages 363–369, July 2012.

[237] Sylvie Despres, Jérôme Nobécourt, and Fanny Rigour. Des primitives visuelles pour l'assistance aux échanges entre experts et ontologues. In Actes de la conférence Ingénierie des Connaissances, Montpellier, France, June 2016.

[238] Rose Dieng, Olivier Corby, Fabien Gandon, Alain Giboin, Joanna Golebiowska, Nada Matta, and Myriam Ribière. Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du knowledge management. Informatiques - Série Systèmes d'information. Dunod, 2001.

[239] Marcos Da Silveira, Julio Cesar Dos Reis, and Cédric Pruski. Management of dynamic biomedical terminologies: current status and future challenges. Yearbook of Medical Informatics, 10(1):125–133, August 2015.

[240] Lucy J.H. Alderson, Phil Alderson, and Toni Tan. Median life span of a cohort of national institute for health and care excellence clinical guidelines was about 60 months. Journal of Clinical Epidemiology, 67(1):52–55, 2014.

[241] Astrid Duque-Ramos, Manuel Quesada Martinez, Miguela Iniesta-Moreno, Jesualdo Tomás Fernández-Breis, and Robert Stevens. Supporting the analysis of ontology evolution processes through the combination of static and dynamic scaling functions in OQuaRE. Journal of biomedical semantics, 7(1):63–63, October 2016.

[242] Anis Tissaoui, Nathalie Aussenac-Gilles, Philippe Laublet, and Nathalie Hernandez. EvOnto : un outil d'évolution de ressource termino-ontologique pour l'annotation sémantique. Technique et Science Informatiques, 32(7-8):817–840, 2013.

[243] Tania Tudorache, Sean Falconer, Csongor Nyulas, Natalya F. Noy, and Mark A. Musen. Will semantic web technologies work for the development of ICD-11? In Proceedings of the 9th International Semantic Web Conference, pages 257–272, Berlin, Heidelberg, 2010. Springer-Verlag.

[244] Tania Tudorache, Csongor I. Nyulas, Natalya F. Noy, and Mark A. Musen. Using semantic web in ICD-11: three years down the road. In Proceedings of the 12th International Semantic Web Conference, pages 195–211, New York, NY, USA, 2013. Springer-Verlag New York, Inc.

[245] Maxim Topaz, Leah Shafran-Topaz, and Kathryn H Bowles. ICD-9 to ICD-10: evolution, revolution, and current debates in the United States. Perspectives in health information management, 10(Spring):1d–1d, April 2013.

[246] Janez Brank, Dunja Madenic, and Marko Groblenik. Gold standard based ontology evaluation using instance assignment. In Proceedings of the 4th Workshop on Evaluating Ontologies for the Web (EON2006), May 2006.

[247] Benjamin M. Good, Gavin Ha, Chi K. Ho, and Mark D. Wilkinson. OntoLoki: an automatic, instance-based method for the evaluation of biological ontologies on the semantic web. CoRR, abs/1502.06025, 2015.

[248] Hao Liu, Ling Zheng, Yehoshua Perl, James Geller, and Gai Elhanan. Can a convolutional neural network support auditing of NCI thesaurus neoplasm concepts? In Proceedings of the 9th International Conference on Biological Ontology, 2018.

[249] Imme Ebert-Uphoff and Yolanda Gil. Exploring synergies between machine learning and knowledge representation to capture scientific knowledge. In Proceedings of the 1st International Workshop on Capturing Scientific Knowledge. Palisades, NY, 2015.

[250] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.

[251] Saïd Abdeddaïma, Sylvestre Vimard, and Lina F. Soualmia. The MeSH-gram neural network model: Extending word embedding vectors with MeSH concepts for semantic similarity. Studies in Health Technology and Informatics, 2019. to appear.

[252] Vimla Patel, Edward H. Shortliffe, Mario Stefanelli, Peter Szolovits, Michael R. Berthold, Riccardo Bellazzi, and Ameen Abu-Hanna. The coming of age of artificial intelligence in medicine. Artificial Intelligence in Medicine, 46(1):5–17, 5 2009.

GLOSSAIRE

ANR	Agence Nationale de la Recherche
BFO	Basic Formal Ontolology
ChEBI	Chemical Entities of Biological Interest
CIM	Classification statistique Internationale des Maladies
CIM-O	Classification statistique Internationale des Maladies pour l'Oncologie
CNHIM	Centre National Hospitalier d'Information sur le Médicament
CUI	Concept Unique Identifier
DDI	Drug-Drug Interaction
DIDEO	Drug-drug Interaction and Drug-drug Interaction Evidence Ontology
DINTO	Drug-Drug Interactions Ontology
DIO	Drug Interaction Ontology
ERIAS	Equipe de Recherche en Informatique Appliquée à la Santé
FIDEO	Food Interacting with Drug Evidence Ontology
FoodOn	Food Ontology
GO	Gene Ontology
IA	Intelligence Artificielle
INO	Interaction Network Ontology
MedDRA	Medical Dictionary for Regulatory Activities
MIAM	Maladies, Interactions Aliments-Médicaments
MSG	Multi-Semantic Group
NCI	National Cancer Institute
NCIt	National Cancer Institute thesaurus
NLM	National Library of Medicine
OAEI	Ontology Alignment Evaluation Initiative
OBO	Open Biomedical Ontologies
OWL	Web Ontology Language
RDF	Resource Description Framework
RTO	Ressource Termino-Ontologique
SemBiP	Semantic BiblioDem Portal
SKOS	Simple Knowledge Organization System
SNCTmt	SNOMED CT mapping tables
SNMI	Systematized NOmenclature of MEDicine - International
SNOMED CT	Systematized NOmenclature of MEDicine - Clinical Terms
SPARQL	SPARQL Protocol and RDF Query Language
UMLS	Unified Medical Language System
W3C	World Wide Web Consortium