Système de gestion lexicale des ressources termino-ontologiques

Guillaume Verdy

Master Sciences, Technologies, Santé
Mention Santé Publique

Spécialité systèmes d'information et technologies
Informatiques pour la santé

Promotion 2021-2022

Système de gestion lexicale des ressources termino-ontologiques

Mémoire réalisé dans le cadre d'une mission effectuée
du 01/03/2022 au 04/09/2022

CHU de Bordeaux
Place Amélie Raba Léon
33000 BORDEAUX

Maitre de Stage :
Romain GRIFFIER, PH
Vianney JOUHET, PH
Bertrand MOAL, Dr

Soutenu publiquement le 05/09/2022
par VERDY Guillaume

Jury de soutenance
Sebastien COSSIN, tuteur universitaire
Gayo DIALLO, rapporteur

Remerciements

À mes trois encadrants : Romain GRIFFIER, Vianney JOUHET et Bertrand MOAL. Vos remarques et conseils ont toujours été pertinents et resteront appréciés face aux futurs problématiques auxquelles je serai confronté dans ma vie professionnelle. Merci pour la patience et la bienveillance dont vous avez toujours fait preuve à mon égard !

À toute l'équipe d'IAM, pour m'avoir accompagné et formé pendant 18 mois, aidé à affiner mon esprit critique, nourris (petits Q, cannelés au rhum et dunes blanches !), et bien évidemment pour les sorties team building du mardi. Je retiendrai principalement *"Mais à quoi ça sert ?"* et *"Contrôle qualité !"*, merci !

Au corps enseignant du Master 2 SITIS, pour le gain de compétences que vous m'avez apporté et qui ne manquera pas de m'être utile dans mon activité professionnelle.

À mes co-internes de Santé Publique, pour nos débats insensés du midi, nos apéros à la plage et nos soirées. C'est un réel plaisir de partager mon internat avec vous !

À mes amis, de Besançon et d'ailleurs, qui savent faire abstraction de mon manque de communication. J'ai beau donner peu de nouvelles, vous savez que le cœur y est !

À ma famille, à qui je dois tout... Merci pour tout !

À Samuel, pour être présent depuis tant d'années, j'ai une chance inouïe de pouvoir te compter parmi mes proches, merci infiniment !

À Nala, pour son amour inconditionnel et réciproque.

"De toutes les passions, la seule vraiment respectable me parait être la gourmandise."
Guy de Maupassant

Table des matières

Liste des tableaux

Table des figures

Abréviations

CIM-10 – Classification Internationale des Maladies, 10e révision
DEAF – Dictionnaire étymologique de l'ancien français
EDS – Entrepôt de Données de Santé
IAM – Informatique et Archivistique Médical
IRI – Internationalized Resource Identifier
ISPED – Institut de Santé Publique, d'Épidémiologie et de Développement
LEMON – LExicon Model for ONtologies
LLOD – Linguistic Linked Open Data cloud
MeSH – Medical Subject Headings
NEL – Named-Entity Linking, ou Annotation Sémantique
NER – Named-Entity Recognition , ou Recherche d'Entités Nommées
RI – Recherche d'Information
TAL – Traitement Automatique de la Langue
RTO – Ressouce Termino-Ontologique
SKOS – Simple Knowledge Organization System
SNOMED-CT – Systematized Nomenclature of Medicine Clinical Terms
STMT – Sub-Term Mapping Tools
TBX – TermBase eXchang
TELIX – Text Encoding and Linguistic Information eXchange
UMLS – Unified Medical Language System
WOLF – Wordnet Libre du Français
WSD – Word-Sense Disambiguation

1 Introduction

1.1 Cadre et sujet du projet

Ce projet s'est déroulé au sein de l'unité d'Informatique et d'Archivistique Médical (IAM) du CHU de Bordeaux, dans le cadre du Master 2 SITIS (Système d'Information et Technologies Informatiques pour la Santé) de l'ISPED (Institut de Santé Publique, d'Épidémiologie et de Développement). De mars à septembre 2022, je me suis attelé à remplir au mieux l'objectif qui m'avait été fixé de développer un système de gestion lexicale des ressources termino-ontologiques. La finalité souhaitée était d'avoir une structure sur laquelle se baser pour développer des services et méthodes qui puissent s'intégrer dans des processus de traitement automatique de la langue (TAL) pour améliorer la recherche d'information (RI) au sein des documents en texte libre des Entrepôts de Données de Santé (EDS), en l'occurrence l'EDS du CHU de Bordeaux.

1.2 Présentation du plan

Afin de structurer au mieux ce mémoire, il a été écrit selon le plan suivant : premièrement, une présentation de la structure m'ayant accueillie, l'unité IAM, expliquant succinctement son rôle et ses missions au sein du CHU. Ensuite, la section *Contexte et justification* permet de mieux comprendre l'intérêt de ce projet et ses objectifs. La section suivante traite des méthodes employées pour la réalisation de ce projet, qui est suivie par la présentation des résultats de ces méthodes employées. Dans la section *Discussion* se trouve les interprétations des résultats précédemment présentés. La section *Conclusion* permet de faire le point sur ce projet, de le critiquer et d'y présenter les possibles perspectives à suivre. Vous pourrez trouver dans la section *Annexes* les différentes définitions, notamment dans le domaine de la linguistique, qui ont été utiles pour traiter au mieux ce projet.

2 Structure d'accueil

L'unité IAM est une unité du CHU de Bordeaux dont le responsable est le Dr Moufid HAJJAR. Elle fait partie du Service d'Information Médicale, lui-même faisant partie du Pôle de Santé Publique du CHU.

L'unité IAM a plusieurs missions. La première consiste en la gestion des archives médicales, la seconde porte sur la gestion des identités des venues à l'hôpital et la troisième concerne l'utilisation secondaire des données de santé, en lien avec l'EDS. Mon stage s'est déroulé en lien avec cette dernière mission.

3 Contexte et justification

3.1 Utilisation secondaire des données de santé

L'utilisation secondaire des données de santé est définie par toute exploitation des données de santé autre que pour leurs usages liés aux soins, qui est l'usage initial de ces données [1, 2]. L'utilisation secondaire des données a de nombreuses finalités, allant de l'évaluation médico-économique d'un établissement de santé à la réalisation d'étude épidémiologique, en passant par le développement d'algorithmes d'intelligence artificielle pour le domaine médical [3]. Ces données peuvent provenir de différentes sources de par l'utilisation de différents logiciels métiers au sein des établissements de santé [4]. Cette hétérogénéité des sources et des données limite leurs utilisations secondaires.

3.2 Entrepôt de Données de Santé

La mise en place d'EDS, nécessitant une étape d'extraction, de transformation et de chargement de ces données organisées en silo, permet de pallier ce frein à l'exploitation secondaire des données de santé [2].

Le CHU de Bordeaux possède un EDS opérationnel depuis 2020. Il contient à ce jour des données sur plus de 2 millions de patients, représentant plus de 2 milliards d'observations. Cet entrepôt a notamment été utilisé pour la production d'indicateurs permettant le suivi de l'épidémie COVID-19 et a également permis au CHU de participer au consortium international 4CE [5], ayant pour but le partage de données de santé agrégées afin d'effectuer des recherches sur l'épidémie de COVID-19. Ce consortium a permis à de nombreuses recherches d'être effectuées avec l'implication du CHU de Bordeaux [6, 7, 8, 9].

Une grande partie des informations médicales contenue dans les EDS hospitaliers [2] est constituée de documents sous forme de texte libre. Ces données, qui sont définies comme étant de type ouvert car sous forme de texte libre, sont opposées à des données que l'on qualifie de type fermé, qui peuvent être par exemple les codes diagnostics attribués au séjour d'un patient.

Les données de type ouvert ne nécessitent pas de processus d'alignement vers un standard (on parle de mapping) préalable à leurs stockages dans un EDS, ce qui est parfois le cas pour les données de type fermé.

Les méthodes et processus appliquées sur les données ouvertes sont donc de nature abstraites et extrapolables aux autres données ouvertes. La généralisation des méthodes d'exploitation de ces données de type ouvert dispose d'une garantie plus sûre qu'avec des données de type fermé. En revanche, extraire les informations contenues dans les données ouvertes nécessitent des processus capable de traiter le texte libre afin d'y repérer les éléments recherchés [10].

3.3 Recherche d'informations

La recherche d'informations (RI) permet de catégoriser des documents comme pertinents ou non au regard d'une information recherchée [11]. Afin que la RI soit efficace sur des données ouvertes, il est utile de mettre en place des processus de TAL [10].

Le TAL permet de traiter des documents non structurés, du texte libre, afin d'y faire de la RI [12]. Le TAL est entre autre limité par l'ambiguïté de certains termes utilisés pour retrouver des documents pertinents [13, 14], mais également par le manque de termes au sein du lexique [1] utilisé [15] .

Le TAL peut se décliner en deux processus que sont la recherche d'entités nommées (NER) et l'annotation sémantique (NEL) [16]. Le NER permet de trouver les termes dans un texte [17] alors que le NEL permet de relier ces termes à des concepts et ainsi leur donner un sens [18].

Si détecter le terme *IVG* dans un document nécessite d'avoir un processus de NER capable de détecter ce terme, le processus de NEL, lui, permet de lier ce terme au concept d'*insuffisance ventriculaire gauche* ou d'*interruption volontaire de grossesse* selon le contexte.

Le TAL permet donc à la fois de détecter les termes et de leur associer les concepts auxquels ils sont reliés. Ainsi, on peut donner un sens aux données contenues dans un documents en texte libre, permettant leurs exploitations.

Deux métriques, le rappel et la précision, permettent d'évaluer l'efficacité d'un processus de RI. La figure 1 est une représentation de ces deux métriques.

Le rappel, appelé également sensibilité, est une métrique permettant d'évaluer l'efficacité de la RI. Il correspond à la mesure du rapport entre le nombre de documents pertinents rapportés par la RI et le nombre total de documents pertinents présents dans le corpus de documents.

La précision, appelée également valeur prédictive positive, est une seconde métrique permettant d'évaluer l'efficacité de la RI. Elle correspond à la mesure du rapport entre le nombre de documents pertinents rapportés par la RI et le nombre total de documents rapportés par la RI.

1. Un lexique est un recueil de termes utilisés dans un domaine. Une définition plus complète est disponible en *Annexes*

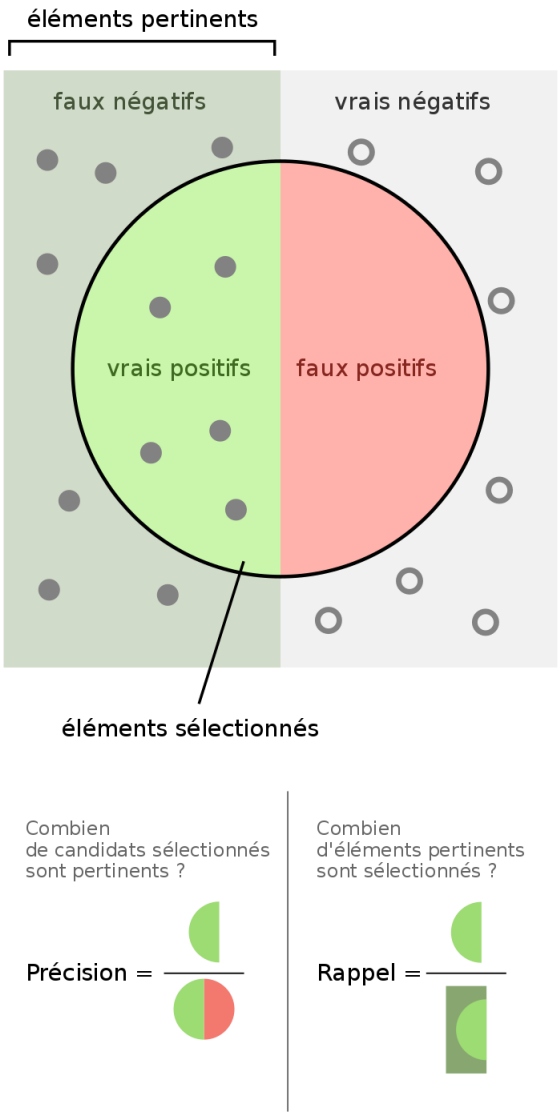

FIGURE 1 – Représentation de la précision et du rappel dans le cadre de la RI
Source: http://fr.wikipedia.org/wiki/Précision_et_rappel

Le Word-Sense Disambiguisation (WSD) permet d'améliorer les métriques de la RI utilisant des méthodes de TAL [19]. Le WSD utilise NER et NEL pour désambiguïser des documents. On peut en déduire qu'optimiser ces deux processus permet d'améliorer les métriques de RI.

En augmentant la richesse des termes d'un lexique utilisé par un processus de NER, on augmente le nombre de termes candidats, et donc le nombre de documents retournés par la RI. Enrichir le lexique utilisé permet donc d'améliorer le rappel mais augmente le risque d'utiliser des termes polysémiques [2], et donc le risque que la

2. Un terme polysémique, ou ambigu, et un terme pouvant signifier plusieurs concepts. Plus de détails en *Annexes*

RI retrouve des documents non pertinents car ambigus. Pour améliorer la précision, détecter ces termes polysémiques est la première étape permettant d'améliorer les processus de NEL [20, 21]. L'enrichissement lexical et la détection des termes polysémiques permet donc d'améliorer la RI portant sur des documents sous formes de texte libre.

3.4 Ressources termino-ontologiques

Les systèmes de représentations de connaissances existent sous de nombreuses formes différentes [22], qui seront désignées sous un terme générique qu'est ressources termino-ontologiques (RTO) [23] dans la suite du document. La figure 2 permet de représenter les relations entre différents systèmes de représentations de connaissances. Une RTO peut se décomposer en deux parties, une partie conceptuelle et un lexique :

- La partie conceptuelle correspond à l'ensemble des concepts, c'est à dire les signifiés des termes. Ce sont les notions représentées dans la RTO. Il existe un standard recommandé pour la gestion de cette partie conceptuelle, le Simple Knowledge Organization System (SKOS) [24].
- Un lexique [25] correspond à l'ensemble des mots, le vocabulaire, utilisé dans une langue ou un domaine de connaissances. Le lexique d'une RTO est donc l'ensemble des termes le constituant.

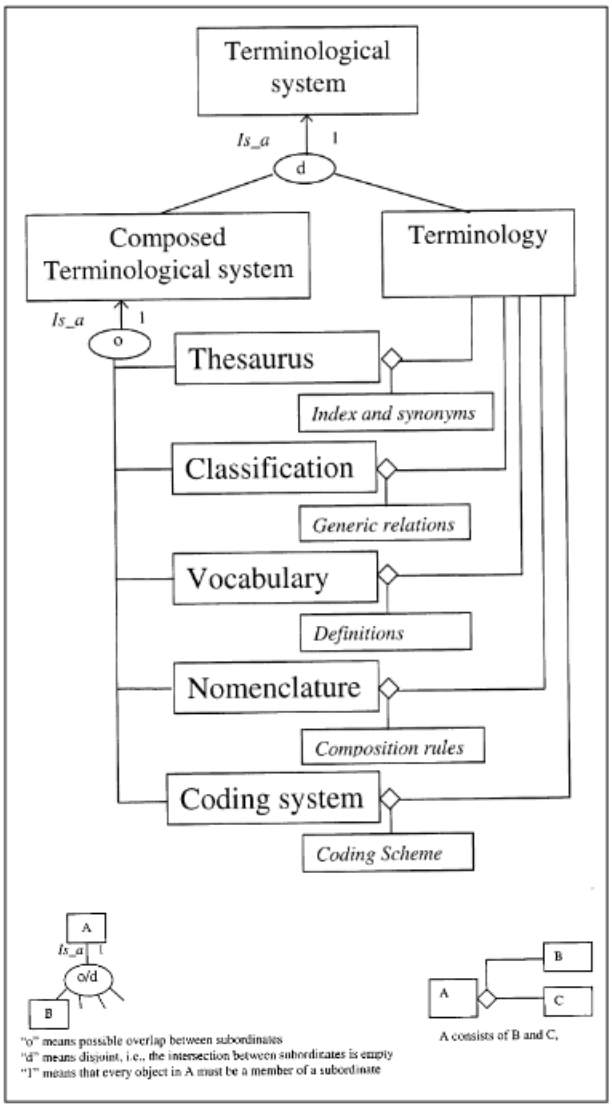

FIGURE 2 – Représentation des relations entre différents systèmes de représentations de connaissances
Source: Understanding Terminological Systems I : Terminology and Typology [26]

3.5 RTO et TAL

Les RTO sont des ressources mettant à disposition les éléments nécessaires aux precessus de TAL à travers leurs lexiques, leurs relations liant leurs lexiques à leurs domaines conceptuels, ainsi que leurs relations entre les concepts eux-mêmes. Les RTO sont donc des ressources utilisées pour les processus de TAL [27, 28].

Les lexiques des RTO ne contiennent pas l'entièreté des termes d'un domaine. On peut alors parler de *mots inconnus*, et il est important des les prendre en considération durant des processus de TAL, puisqu'ils sont une limite à la RI [29].

3.6 Problématique

SKOS est un modèle permettant de gérer la partie conceptelle des RTO, en fournissant le vocabulaire permettant de décrire les relations entre les concepts. SKOS propose également un vocabulaire pour décrire les relations entre les concepts et les termes d'un lexique : skos:prefLabel, skos:altLabel et skos:hiddenLabel. Ces relations sont toutes des sous-propriétés de la relation rdfs:label, et permettent de décrire à des degrés divers la relation entre un concept et un terme.

En revanche, SKOS n'est pas un modèle adapté pour traiter de la partie lexicale des RTO [30]. Ce modèle ne fournit pas de vocabulaires permettant de décrire la décomposition d'un terme, de le décrire d'un point de vue linguistique, en donnant les différentes relations entre les formes d'un lemme [3], de donner sa classe grammaticale, ou encore de fournir des informations syntaxiques. SKOS n'est donc pas la solution pour le développement d'une structure de gestion de la partie lexicale des RTO dans le but de développer des processus de TAL. Or, avoir un modèle de gestion lexicale de ces RTO permettrait d'améliorer les processus de TAL [31], en proposant le développement d'outils d'enrichissement lexical ou de détections de termes ambigus par exemple.

SKOS ne permet pas, par exemple, de déterminer que le terme *défaillance ventriculaire gauche* et composé du terme *défaillance*, qui est lui-même synonyme du terme *insuffisance*, et en conclure que *défaillance ventriculaire gauche* est synonyme du terme *insuffisance ventriculaire gauche*, enrichissant ainsi la base de donnée lexicale permettant d'améliorer les processus de NER, et donc le rappel d'une RI.

SKOS ne permet par non plus de décrire les cadres syntaxo-sémantiques d'un terme. Si l'on prend l'exemple du terme *inflammation*[32]. Ce terme peut rapporter au concept d'*inflammation systémique* ou d'*inflammation localisée*. Son sens peut être déterminé en fonction de son cadre syntaxo-sémantique au sein d'une phrase. *L'inflammation a augmenté de 5 cm* aura pour signifié l'*inflammation localisée*, car une notion de mesure y est associée. La détermination du concept associé est déterminé par le cadre syntaxo-sémantique du terme.

Les processus de TAL, et donc par extension les processus de RI sur des documents en texte libre, pourraient gagner en performance via l'utilisation de services et méthodes se basant sur une structure de gestion lexicale des RTO.

3.7 Objectif

Un modèle de gestion lexicale permettrait de développer des services et méthodes abstraits et donc réutilisable pour améliorer les performances de RI en fournissant des outils d'aides aux processus de TAL. Il n'apparaît par que l'implémentation de structure de gestion lexicale, permettant d'aider les processus de TAL, soit aujourd'hui facilement disponible.

3. Un lemme est une entrée de lexique. Les formes peuvent être par exemple le pluriel d'un nom ou le féminin d'un adjectif.

L'objectif de ce travail est donc d'identifier un modèle de gestion lexicale, de l'implémenter et d'y développer des services et méthodes. Ces services et méthodes ont pour but de contribuer à aider et améliorer des processus de TAL et ainsi perfectionner des RI portant sur des documents médicaux peu structurés, tels que ceux stockés dans des EDS, notamment l'EDS du CHU de Bordeaux.

4 Méthode

4.1 Choix d'un modèle de gestion lexicale

La première étape a consisté à trouver un modèle permettant la description de lexiques. Le choix du modèle s'est porté sur celui répondant le mieux aux critères suivants :

- Le modèle doit être implémentable selon un schéma RDF. En effet, SKOS est une recommandation de la W3C permettant la gestion de la partie conceptuelle d'un lexique, ce qui est nécessaire dans les processus de TAL. Ce standard étant basé sur un format RDF, le modèle de gestion lexicale doit l'être également. Les deux modèles étant complémentaires, l'utilisation d'un même format RDF facilite leur inter-opérabilité.
- Le modèle doit permettre de faire les liens avec le modèle SKOS, toujours dans le but d'une inter-opérabilité entre les deux modèles.
- Le modèle doit pouvoir répondre aux besoins de gestions lexicales auxquels SKOS n'est pas adapté afin de permettre à des méthodes de TAL de se rapprocher de la compréhension humaine [30]. Le modèle doit permettre de nuancer les concepts signifiés par une entrée lexicale, de mieux décrire les attributs d'une entrée lexicale, tel que sa classe grammaticale, et également de renseigner des informations sur des règles syntaxiques liées au sens d'une entrée lexicale. Cela permet d'assurer que le modèle réponde à nos attentes, en permettant d'exploiter efficacement des processus d'enrichissement lexical, de détection de polysémie ou de synonymie entre les termes du lexique.
- Le modèle doit être suffisamment abstrait, permettant le développement de méthodes génériques qui puissent être applicables à des situations diverses. Il ne doit donc pas être centré sur une langue ou un domaine de connaissance.
- Le modèle doit être publié dans une version finale. Cela permet d'avoir une meilleure garantie de la qualité du modèle, nécessaire pour y développer des services devant fonctionner de manière abstraite et stable.
- Le modèle doit être un standard recommandé de la W3C, permettant de garantir une meilleure inter-opérabilité avec d'autres standards recommandés.

Afin de trouver un modèle pouvant correspondre à ces critères, des recherches sur la plateforme de recherche google scholar. Les requêtes effectuées ayant permis

de trouver des modèles pertinents sont les suivantes : *linguistic rdf model*, *linguistic semantic resources rdf*, *NLP Lexicon model* et *NLP Lexicon tool*.

Une fois le modèle sélectionné, une présentation détaillée des parties pertinentes pour la mise en place de la structure de gestion lexicale sera effectuée.

Le modèle choisit devant également montrer son efficacité pour répondre aux problèmes soulevés par les termes synonymes et polysémiques, il a été choisi pour illustrer de modéliser au sein de ce modèle cinq termes permettant d'appréhender cette problématique, que sont *Insuffisance ventriculaire gauche*, *Insuffisances ventriculaires gauches*, *Interruption volontaire de grossesse*, *Interruptions volontaires de grossesses* et *IVG*. En effet, *IVG* est polysémique car doté de plusieurs sens, *Insuffisance ventriculaire gauche* et *Interruption volontaire de grossesse* sont deux de ses synonymes.

4.2 Implémentation de la structure de gestion lexicale

Une fois le modèle de gestion lexicale choisit, il est nécessaire de l'implémenter en une structure de gestion lexicale. Les outils techniques utilisés pour son développement sont sommairement décris, ainsi que les méthodes employées afin de contrôler la conformité de l'implémentation en regard du modèle choisi.

4.2.1 Outils utilisés pour l'implémentation de la structure

RDF4J [33] est une infrastructure logicielle en libre accès permettant de manipuler des données RDF dans le langage informatique java et a donc été utilisé pour implémenter le modèle. Cette infrastructure fournit de nombreux outils permettant entre autre de générer des modèles RDF et de faciliter la communication avec une base de données orientée graphe.

L'équipe encadrant ce projet a déjà développé des outils permettant des utilisations secondaires des données de santé se basant sur une architecture de type micro-service. Cette architecture a été développée à l'aide d'un cadre de développement, Jhipster [34]. Puisque cette structure gestion lexicale a pour vocation d'être implémenté au sein de cette architecture, Jhipster a également été utilisé pour son élaboration.

4.2.2 Transformation de RTO vers le modèle adéquat

Le but de cette structure étant de prendre en charge la partie lexicale des RTO, une étape de transformation de RTO vers le modèle sélectionné est nécessaire. Pour cela, les différentes classes manipulées dans la structure sont décrites, ainsi que l'explication du fonctionnement de l'algorithme permettant d'en générer des instances à partir des données contenues des RTO.

4.2.3 Contrôle de la structure implémentée

Une fois la structure implémentée, il est nécessaire de s'assurer de sa justesse. Le nombre d'éléments de chaque classe ainsi que les cardinalités des relations entre ces éléments doivent être cohérents en regard du modèle sélectionné. Des requêtes SPARQL sont effectuées à cet effet.

Puisque l'on souhaite que le modèle choisit puisse traiter efficacement les problèmes de polysémie et de synonymie, il a été incorporé dans la structure les 5 termes énoncés précédemment permettant d'illustrer ces notions. Ces 5 termes sont : *Insuffisance ventriculaire gauche*, *Insuffisances ventriculaires gauches*, *Interruption volontaire de grossesse*, *Interruptions volontaires de grossesses* et *IVG*.

Il a également fallu contrôler la robustesse de la structure face à des données plus riches. Pour cela, il a été choisit de charger dans la structure quatre RTO contenant de nombreuses informations : la Medical Subject Headings (MeSH), la Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) la Classification Internationale des Maladies, 10e révision (CIM-10) et ORPHANET. Des métriques permettant leurs transformations dans le modèle choisi sont présentées.

4.3 Développement de services et méthodes

Une fois la structure implémentée et que ses performances évaluées, des services et méthodes ont été développés. Certains sont d'un intérêt pratique de gestion de cette structure, d'autres afin de répondre à des besoins pratiques en matière de NLP, illustrées par des méthodes fournissant des métriques pertinentes sur les notions polysémie et de synonymie.

5 Résultats

5.1 Choix du modèle de gestion lexical

Le premier modèle ayant été analysé est TELIX [35] (Text Encoding and Linguistic Information eXchange). Ce modèle RDF a pour objectif de permettre l'annotation d'éléments linguistiques. Il se construit sur le vocabulaire SKOS-XL [24], qui est l'extension de SKOS permettant de décrire les liens entre les concepts et les entrées lexicales. TELIX permet par exemple de définir par une annotation que deux entrées lexicales sont synonymes.

Le second modèle analysé est le Specialist Lexicon [36]. Centré sur l'Unified Medical Language System (UMLS) [37], il est avant tout une base de données lexicales fonctionnant de pair avec le Lexical Tools [38], un service proposant des outils de gestion lexicale, comme le Sub-Term Mapping Tools (STMT) [39]. Le STMT permet par exemple un découpage en sous-termes des entrées lexicales retrouvées dans un

lexique, pouvant permettre la création de nouveaux synonymes. Ce modèle n'est pas directement disponible au format RDF, et ne permet pas un niveau d'abstraction suffisant. Certaines classes du modèle sont par exemple construites autour d'un seul langage, l'anglais.

D'autres bases de données lexicales ont été analysées, telles que WordNet [40] et DBpedia [41]. Ces structures, comme le Specialist Lexicon, présentent le désavantage d'avoir été constituées avec un niveau d'abstraction linguistique insuffisant, souvent centré initialement en anglais, et devant répondre à une problématique d'un domaine en particulier. Leurs modèles de construction sont difficlement accessibles, ce qui limite leurs utilisations dans la création d'une structure de gestion lexicale.

Le troisième modèle ayant été analysé est Lemon [42, 43, 44](LExicon Model for ONtologies). Ce modèle RDF permet de décrire les entrées lexicales d'un lexique tout en décrivant leurs relations avec des concepts. Il est composé d'un core autonome, de modules complémentaires, et utilise LexInfo [30], un modèle RDF fournissant un vocabulaire linguistique.

Le quatrième modèle analysé est OntoLex-Lemon [45], la version définitive de Lemon. Ontolex est un consortium regroupant des participants de domaines divers travaillant sur l'élaboration de modèles de représentation linguistique depuis 2011. Plusieurs de ces modèles ont été combinés pour former le modèle Ontolex-Lemon. Ce modèle a été publié en 2016 par la W3C, et est donc celui qui a été retenu comme modèle servant à l'implémentation de la structure, puisqu'il répond le mieux aux critères de sélections énoncés précédemment.

La table 1 permet de visualiser la validation des critères de sélections selon les différents modèles analysés.

Table 1 – Tableau représentant les différents modèles de gestion lexicale analysés selon les critères de pertinences
Source: Représentation basée sur les critères de pertinences de la partie 4.1

	TELIX	Specialist Lexicon	Lemon	Ontolex-Lemon
Modèle RDF	✓	✗	✓	✓
Relations avec modèle SKOS	✓	✗	✓	✓
Description linguistique suffisante	✗	✓	✓	✓
Niveau d'abstraction suffisant	✓	✗	✓	✓
Version finale	✓	✓	✗	✓
Standard W3C	✗	✗	✗	✗

5.2 Ontolex-Lemon

Ontolex-Lemon est un modèle fournissant un vocabulaire basé sur un format RDF utilisant un maximum de standard de la W3C dans ce domaine, mais n'est pas un standard recommandé par la W3C lui-même. Il permet de décrire la grammaire d'un lexique, tant du point de vue syntaxique, morphologique que sémantique. Il est donc utilisable pour décrire la partie lexicale des RTO, et constituer une base de données orientée graphe depuis le lexique de celles-ci. [46] [42]. Il utilise comme source de vocabulaires permettant de donner des informations lexicales LexInfo.

Ce modèle est doté d'un Core, qui décrit les principales classes du modèle, et de modules complémentaires, décrivant des classes et relations supplémentaires pour des besoins linguistiques plus poussés. La figure 3 est une représentation des classes du Core d'Ontolex-Lemon ainsi que des relations les liant. Les instances des classes d'Ontolex-Lemon ont chacune une Internationalized Resource Identifier (IRI), permettant de les identifier individuellement dans une base de données orientée graphe.

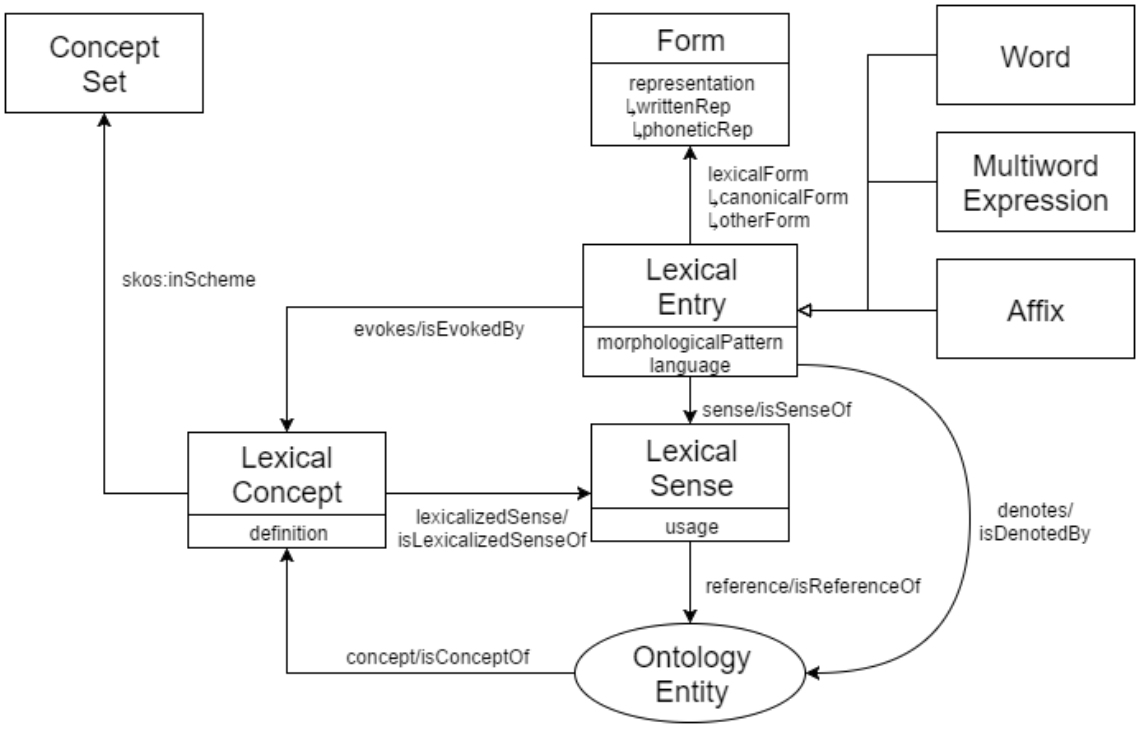

FIGURE 3 – Représentation du Core d'Ontolex-Lemon
Source: https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

5.2.1 Principales classes

Les paragraphes suivants présentent successivement les différentes classes retrouvées au sein du Core pertinentes à la mise en place d'une structure de gestion lexicale que sont : ontolex:LexicalEntry, ontolex:Form, ontolex:LexicalSense, ontolex:OntologyEntity. Une cinquième classe sera décrite, skos:ConceptScheme, car elle

est décrite dans le modèle Ontolex-Lemon et pertinente pour la conception de la structure, mais provient du modèle SKOS. Les noms de domaines des classes précédemment énoncés ne sont pas réécris dans la suite du document pour une meilleur lisibilité.

LexicalEntry : c'est la classe centrale du modèle. Elle consiste en un ensemble de Form cohérent au regard d'un ensemble de LexicalSense. Une LexicalEntry peut être considérée comme un lemme, dans le sens d'entrée de lexique.

Elle est reliée à

- Un ensemble de Form : au minima une canonique, et potentiellement à d'autres qui sont les formes fléchies de la canonique. Il existe donc trois relations possibles d'une LexicalEntry à une Form : ontolex:lexicalForm, ontolex:canonicalForm et ontolex:otherForm. Ces deux dernières relations sont des sous-relations de la première.
- Un ensemble de LexicalSense. Une LexicalEntry est liée à un LexicalSense par la relation ontolex:sense.
- Un ensemble d'OntologyEntity. Une LexicalEntry est liée à une OntologyEntity par la relation ontolex:denotes.

Form : c'est la classe permettant de représenter une réalisation grammaticale d'une instance de LexicalEntry. La représentation écrite de la classe Form se fait au travers de la relation ontolex:writtenRep vers un rdf:langString.

LexicalSense : une instance de LexicalSense n'est reliée qu'à une unique LexicalEntry par la relation ontolex:isLexicalSenseOf, et une unique OntologyEntity, par la relation ontolex:reference. Cette classe est principalement utile pour d'autres modules d'Ontolex-Lemon et sert d'interface entre une LexicalEntry et une OntologyEntity, permettant d'ajouter des informations sur cette relation au travers de ce LexicalSense. Elle permet notamment de nuancer la signification d'une LexicalEntry, de faire des liens avec un cadre syntaxique et également d'apporter des métadonnées sur la relation entre LexicalEntry et OntologyEntity.

OntologyEntity : c'est une classe de haut niveau d'abstraction, correspondant à une rdf:Ressource, l'entité englobant toutes les entités du modèle RDF. Elle est liée à un ensemble de LexicalSense et un ensemble de LexicalEntry, par les relations ontolex:isReferenceOf et ontolex:isDenotedBy respectivement. OntologyEntity est représenté dans ces résultats uniquement par la classe skos:Concept, une classe du modèle SKOS, qui est liée à un ConceptScheme par la relation skos:inScheme.

ConceptScheme : un ConceptScheme est l'entité agrégeant un ensemble de Concept appartenant à une RTO, ainsi que les relations entre ces Concept.

5.2.2 Modules complémentaires

Des modules s'articulant autour du Core sont décrit dans le modèle, il ne sera présenté sommairement que les modules Decomp et Synsem, qui apparaissent comme pertinents pour la mise en place de services en lien avec le TAL.

***Decomp* :** Ontolex-Lemon possède un module permettant de décrire la décomposition d'une LexicalEntry : *Decomp*. Elle fait intervenir de nouvelles classes comme Component, et des relations procurant des informations sur la manière donc une LexicalEntry peut être décomposée.

***Synsem* :** Ontolex-Lemon possède un module permettant de décrire les règles syntaxiques d'une LexicalEntry et d'y lier des informations sémantiques : *Synsem*. Elle fait intervenir de nouvelles classes telles que SyntacticFrame ou SyntacticArgument, qui permettent par exemple de définir le domaine d'un COD d'un verbe intransitif et sa position par rapport à ce verbe au sein de la phrase.

5.3 Implémentation d'une structure selon le modèle Ontolex-Lemon

5.3.1 Définition du domaine

Premièrement, il a été nécessaire de définir les éléments du domaine de ce modèle. Le domaine est composé des différentes classes définies par Ontolex-Lemon dont la manipulation est nécessaire pour répondre aux besoins du projet.

Les cinq classes essentielles utilisées pour la structure sont : LexicalEntry, LexicalSense, Form, OntologyEntity et ConceptScheme. OntologyEntity est une classe avec un haut niveau d'abstraction dans le format RDF, puisqu'elle est de la classe rdfs:Ressource. Au sein du projet, son cadre a été limité à une classe du standard SKOS, celle de Concept. Bien définir ces éléments du domaine consiste à s'assurer que chaque instance de ces classes soit cohérente en regard du modèle.

Les instances de ces classes sont identifiées par des IRI. Les IRI de chaque instance doivent être différentes ou identiques selon les définitions du modèle Ontolex-Lemon. Ici sont définit les conditions nécessaires à ce que deux instances d'une même classe possèdent la même IRI, et sont donc identiques.

- Deux instances de Form sont identiques si les ontolex:writtenRep sont les mêmes, c'est à dire si leurs représentations écrites au travers d'un rdf:langString sont les mêmes.
- Deux instances de LexicalEntry sont identiques si les instances de Form de ces LexicalEntry sont les mêmes. Toutes les Form d'une LexicalEntry doivent être cohérentes avec tous les LexicalSense de la LexicalEntry.
- Deux instances de LexicalSense sont identiques s'ils sont issus du même Concept et de la même LexicalEntry.
- La classe Concept est une sous-classe d'OntologyEntity. Un concept d'une RTO est représenté par un code unique au regard de la RTO dont il est issu.

Deux instances de Concept sont donc identiques s'ils partagent le même code et la même RTO.

5.3.2 Algorithme de transformation

Au vu des conditions définies précédemment, un algorithme de création des IRI des cinq classes du domaine à partir de chaque couple concept / terme de la RTO que l'on souhaite transformer au format Ontolex-Lemon a été créé.

Afin de transformer une RTO dans un format conforme au modèle Ontolex-Lemon, cette RTO doit être constituée :

- D'un identifiant unique à la ressource.
- D'un Namespace.
- D'une liste de concepts. Chaque concept est constitué d'une chaîne de caractère alphanumérique correspondant au code du concept.
- D'une liste de termes par concept. Chaque terme est constitué d'une chaîne de caractère alphanumérique.

La figure 4 représente une partie d'une RTO selon le modèle SKOS, la version française 2022 de la MedDRA, à l'aide de deux concepts et de cinq labels, dont *IVG* qui est un label commun aux deux concepts. Il est à noter qu'au sein de cette RTO la relation entre le concept MedDRA:10024119 et *IVG* n'existe pas, mais il a été choisi de la décrire afin de créer un exemple pertinent.

Concept : une instance de la classe Concept correspond à l'IRI formée par la concaténation entre le Namespace de la RTO et de la chaîne de caractère alphanumérique d'un concept. D'après l'exemple, deux instances de la classe Concept serait créées, une pour chaque concept.

Pour les instances des autres classes, le Namespace a été définit comme étant le suivant : *http ://www.chu-bordeaux.fr/lexic-manager#*.

ConceptScheme : Une instance de la classe ConceptScheme correspond simplement à la concaténation du namespace et du nom de la RTO dont sont issus les concepts. D'après l'exemple, une seule instance de la classe ConceptScheme doit exister.

Form : une instance de la classe Form correspond à l'IRI issue de la concaténation du Namespace, de la chaîne de caractère *form*, et du hash du label. D'après l'exemple, cinq instances de la classe Form sont créés, une pour chaque rdfs:label des concepts.

LexicalEntry : une instance de la classe LexicalEntry correspond à l'IRI issue de la concaténation du Namespace, de la chaîne de caractère *lexicalentry*, et du hash de la concaténation de la liste des IRI triée par ordre alphanumérique de l'ensemble des Form d'une LexicalEntry donnée, c'est à dire de sa forme canonique et de ses

potentielles formes fléchies. D'après l'exemple fourni par la figure 4, trois instances de la classe LexicalEntry seraient créées, puisqu'on y trouve trois Form dites canoniques, c'est à dire non fléchies.

LexicalSense : une instance de la classe LexicalSense correspond à la concaténation entre le Namespace, la chaîne de caractère *lexicalsense* et du hash issu de la concaténation de l'IRI de l'instance du Concept auquel il est lié et de l'IRI de l'instance de la LexicalEntry auquel il est lié également. D'après l'exemple, quatre instances de la classe LexicalSense seraient créées, une pour chaque couple instance de Concept / instance de LexicalEntry.

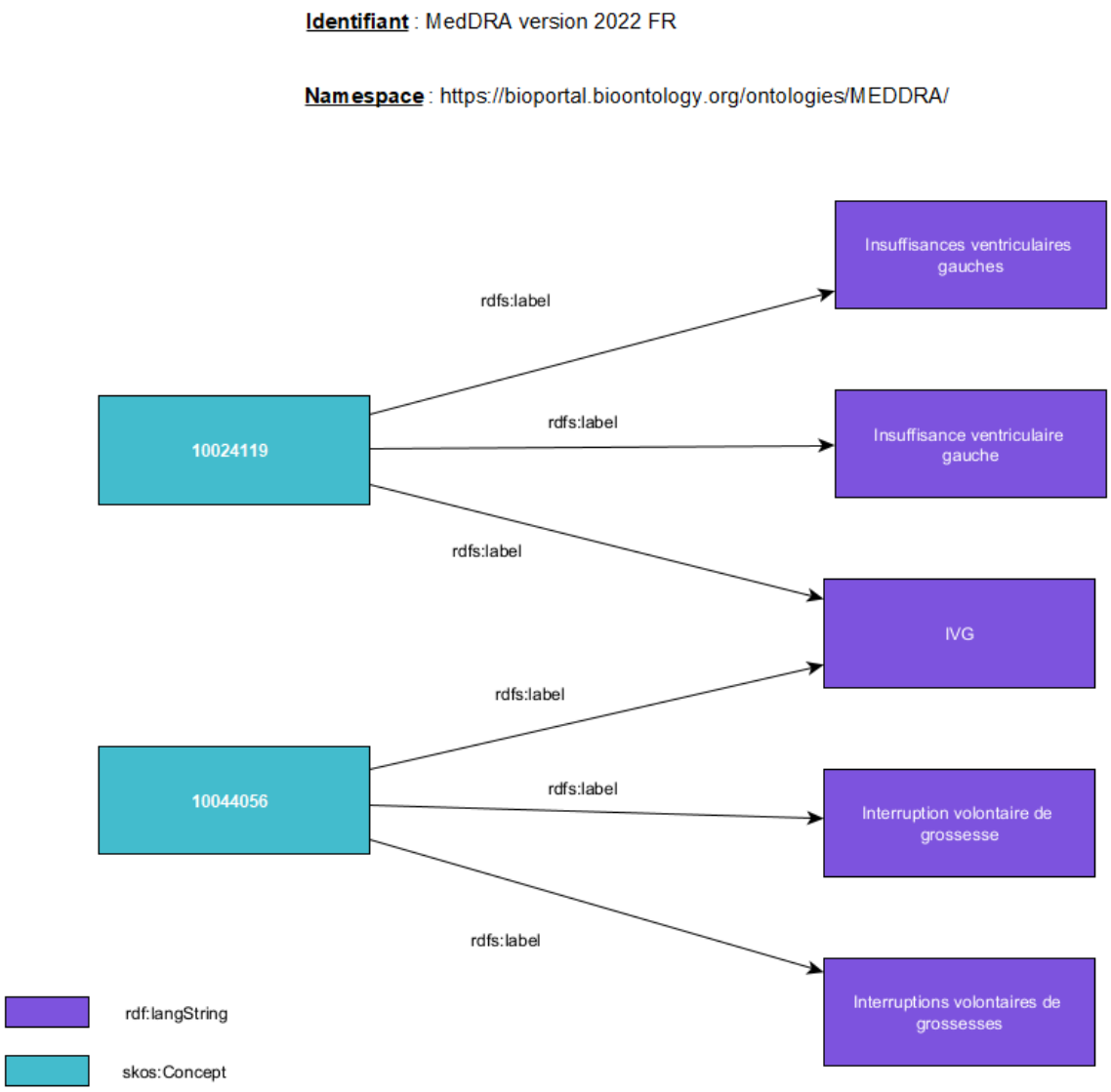

FIGURE 4 – Représentation d'une partie de RTO MedDRA avec les éléments nécessaires à sa transformation dans un format Ontolex-Lemon
Source: Création d'après le modèle SKOS [24]

5.3.3 Constitution des triplets RDF conformes au vocabulaire Ontolex-Lemon

Une fois les IRI des objets du domaine créées, il est nécessaire de créer un modèle RDF constitué de triplets établissant les relations entre ces différentes instances. La figure 5 représente les classes du domaine, certaines relations entre ces classes et les cardinalités de ces relations. Ainsi, selon l'exemple fournis par la figure 4, les relations

suivantes doivent être retrouvées :

- Quatre relations de type ontolex:sense.
- Quatre relations de type ontolex:isSenseOf.
- Quatre relations de type ontolex:reference.
- Quatre relations de type ontolex:isReferenceOf.
- Quatre relations de type ontolex:denotes.
- Quatre relations de type ontolex:isDenotedBy.
- Cinq relations de type ontolex:lexicalForm.
- Trois relations de type ontolex:canonicalForm (sous propriété de la relation ontolex:lexicalForm).
- Cinq relations de type ontolex:writtenRep.

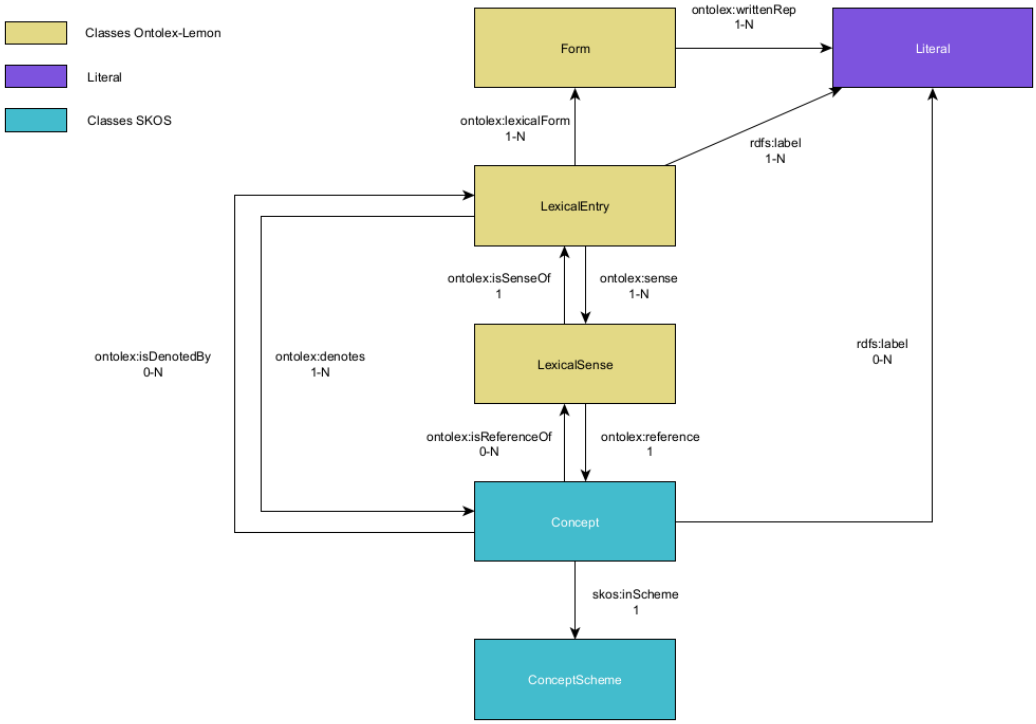

FIGURE 5 – Représentation des classes principales utilisées dans le projet, leurs relations et les cardinalités de ces relations
Source: création d'après le modèle Ontolex-Lemon [45]

Ces triplets RDF peuvent être assignés à des Graphs. L'IRI d'un Graph est donné par la concaténation entre le namespace *http ://www.chu-bordeaux.fr/lexic-manager#* et le hash de l'identifiant de la RTO. Chaque triplet généré est affilé à son Graph correspondant.

5.4 Contrôle de la structure

5.4.1 Règles pour la cohérence du modèle

Des règles se basant sur le nombre d'éléments d'Ontolex-Lemon et les cardinalités des relations entre ces éléments ont été établis :

— Concernant le nombre d'éléments :

- il ne peut pas y a voir plus de LexicalEntry que de Form, puisqu'une LexicalEntry se construit par un ensemble de Form. Il y a donc au minima autant de LexicalEntry que de Form dans le cas où il n'y a qu'une seul Form par LexicalEntry.
- il ne peut pas y a voir moins de LexicalSense que de LexicalEntry ou de Concept, puisqu'un LexicalSense est construit pour chaque couple composé d'un Concept et d'une LexicalEntry.

— Concernant la cardinalité des relations entre les éléments :

- le nombre de ontolex:canonicalForm est inférieur ou égal au nombre de ontolex:lexicalForm, puisque ontolex:canonicalForm et une sous propriété accessoire d ontolex:lexicalForm.
- les relations inverses (ontolex:sense et ontolex:isSenseOf, ontolex:reference et ontolex:isRefereneOf, ontolex:denotes et ontolex:isDenotedBy) doivent au sein de chaque couple être identique, puisque ce sont des relations symétriques.
- les relations issues des éléments LexicalSense doivent toutes être au même nombre. La cardinalité depuis cette classe étant de un pour un pour chaque relation, on doit retrouver autant de ontolex:sense que de ontolex:reference.
- le nombre de ontolex:denotes ne peut pas être supérieur au nombre de ontolex:sense et ontolex:reference.

5.4.2 Évaluation des données générées

Pour contrôler la justesse de la structure, il a été choisi de transformer les données de la figure 4 au travers de l'algorithme développé. La table 2 contient le nombre d'éléments du domaine présent, et la table 3 contient le nombre de relations entre ces éléments présentes dans la base de données graphe.

La figure 6 permet d'avoir une représentation des objets du domaine ainsi qu'une partie des relations entre ces entités construite depuis les données de la figure 4.

TABLE 2 – Tableau représentant le nombre d'éléments par rdfs:Class issus de l'exemple de la figure 4
Source: requête SPARQL de la base de données graphe

rdfs:Class	Nombre d'éléments
ConceptScheme	1
Concept	2
LexicalSense	4
LexicalEntry	3
Form	5

TABLE 3 – Tableau représentant le nombre d'éléments par relations issus de l'exemple de la figure 4
Source: requête SPARQL de la base de données graphe

Predicat	Nombre de relations
ontolex:sense	4
ontolex:isSenseOf	4
ontolex:reference	4
ontolex:isReferenceOf	4
ontolex:denotes	4
ontolex:isDenotedBy	4
ontolex:lexicalForm	5
ontolex:canonicalForm	3
ontolex:writtenRep	5

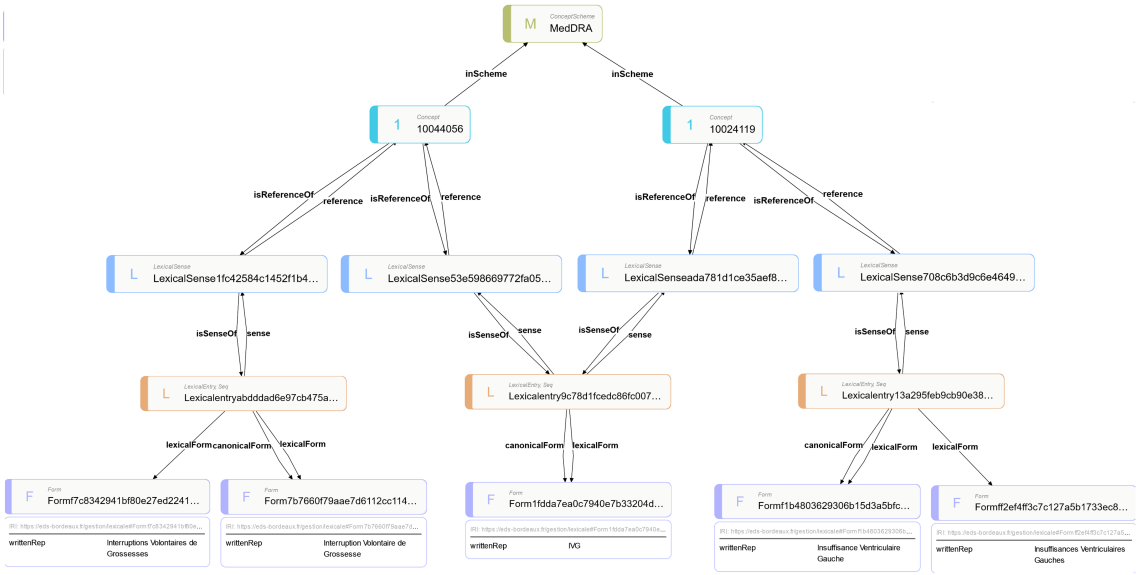

FIGURE 6 – Représentation de la base de données graphes contenant des données de contrôle
Source: données graphes visualisées par l'outil Ondodia

Ensuite, quatre RTO en langue française transformées selon le modèle Ontolex-Lemon ont été chargées dans la base de données graphe. La CIM-10, le MeSH, Orphanet et la SNOMED-CT ont été choisis. Ces RTO ont été sélectionnées car leurs contenus est pertinents, les quatre recensant des données sur des pathologies.

De manière similaire au table 2 et 3, les Table 4 et 5 contiennent respectivement le nombre d'éléments du domaine présent et le nombre de relations entre ces éléments présentes dans la base de données graphe.

TABLE 4 – Tableau représentant le nombre d'éléments par rdfs:Class issu des quatre RTO importées
Source: requête SPARQL de la base de données graphe

rdfs:Class	Nombre d'éléments
ConceptScheme	4
Concept	85 221
LexicalSense	194 939
LexicalEntry	192 977
Form	192 977

TABLE 5 – Tableau représentant le nombre d'éléments par relations issu des quatre RTO importées
Source: requête SPARQL de la base de données graphe

Predicat	Nombre de relations
ontolex:sense	194 939
ontolex:isSenseOf	194 939
ontolex:reference	194 939
ontolex:isReferenceOf	194 939
ontolex:denotes	194 939
ontolex:isDenotedBy	194 939
ontolex:lexicalForm	192 977
ontolex:canonicalForm	192 977
ontolex:writtenRep	192 977

5.5 Développement de services et méthodes

5.5.1 Charger une RTO dans une base de données graphe

Le premier service développé permet de transformer une RTO selon le modèle Ontolex-Lemon et de le charger dans une base de données graphe. C'est ce service qui a été utilisé pour modéliser les quatre RTO présentées dans le contrôle de la structure.

5.5.2 Obtenir la liste des RTO présentes dans une base de données orientée graphe

Le second service implémenté a été la possibilité d'obtenir la liste des RTO présente dans la base de données orientée graphe, avec des métadonnées renseignant sur le contenu des RTO retrouvées. Une requête SPARQL est utilisée afin de recueillir la liste des RTO ainsi que le nombre de LexicalEntry et de Concept présentes dans celle-ci. La table 6 représente le nombre de Concept et de LexicalEntry par ConceptScheme suite à l'importation de ces RTO.

TABLE 6 – Tableau représentant le nombre de Concept et de LexicalEntry par ConceptScheme
Source: requête SPARQL de la base de données graphe

ConceptScheme	Nombre de Concept	Nombre de LexicalEntry
MESH	29 351	122 828
SNOMED-CT	26 162	42 401
CIM-10	19 033	19 018
ORPHANET	10 675	10 674

5.5.3 Termes polysémiques

Une méthode permettant de déterminer le nombre de termes ambigus a été développée. Un terme ambiguë se définit comme une LexicalEntry ayant plus d'une relation ontolex:denotes, c'est à dire une LexicalEntry signifiant plusieurs concepts. La table 7 représente le nombre de LexicalEntry par nombre de Concept liés à ces LexicalEntry.

TABLE 7 – Tableau représentant le nombre de LexicalEntry polysémiques avec l'ensemble des RTO
Source: requête SPARQL de la base de données graphe

Concept par LexicalEntry	Nombre de LexicalEntry (%)
1	191 148 (99,05)
2	1 696 (0.88)
3	133 (0.07)

La table 8 fournit le nombre de LexicalEntry polysémiques par ConceptScheme, c'est à dire avec un nombre de Concept lié supérieur à un.

Individuellement, on retrouve donc 18 LexicalEntry polysémiques en tout pour ces quatre RTO, contre 1 829 lorsque ces RTO sont regroupées.

Parmi les 15 lexicalEntrys polysémiques de la CIM-10, on retrouve notamment le terme *Syphilis cardio-vasculaire*, qui est le signifiant d'après notre modèle de deux concepts, le A520 et le I980, ou encore *Asphyxie*, pour les concepts R090 et T71.

TABLE 8 – Tableau représentant le nombre de LexicalEntry polysémiques par ConceptScheme
Source: requête SPARQL de la base de données graphe

ConceptScheme	Nombre de LexicalEntry polysémiques
MESH	2
SNOMED-CT	0
CIM-10	15
ORPHANET	1

Certaines lexicalEntrys ont des relations avec des concepts issus de RTO différentes, tel que le terme *Syndrome de Lynch*, qui est lié au concept C189+0 de la CIM-10, 144 d'OrphaNet et D003123 du MeSH.

5.5.4 Termes synonymes

Une méthode permettant de déterminer le nombre de termes synonymes a été développée. Des termes sont dit synonymes s'ils sont liés à un même concept par la relation ontolex:denotes. La table 5.5.4 représente le nombre de Concept par nombre de LexicalEntry liés à ces Concept.

TABLE 9 – Tableau représentant le nombre de Concept par le nombre de LexicalEntry synonymes
Source: requête SPARQL de la base de données graphe

LexicalEntry par Concept	Nombre de Concept (%)
1	52 239 (61,30)
2	13 036 (15,30)
3	5 826 (6,85)
4	4 349 (5,10)
5	2 500 (2,93)
6	1 999 (2,35)
7	1 234 (1,45)
8	957 (1,12)
9	684 (0,80)
10	506 (0,59)
Plus de 10	1 891 (2,21)

6 Discussion

6.1 Choix d'Ontolex-Lemon

Le choix d'utiliser Ontolex-Lemon comme modèle pour développer une structure de gestion lexicale se base sur des critères cherchant à s'assurer de la pérennité du projet. En sélectionnant un modèle recommandé par le W3C et conçut par des experts du domaine de la linguistique, il est espéré que la structure permette d'être utilisée de façon efficace pour des processus de RI au sein de l'EDS du CHU de Bordeaux. Bien que recommandé par la W3C, Ontolex-Lemon n'a pas le statut de standard de ce domaine, mais il est le modèle ayant le mieux répondu aux critères de sélections établis.

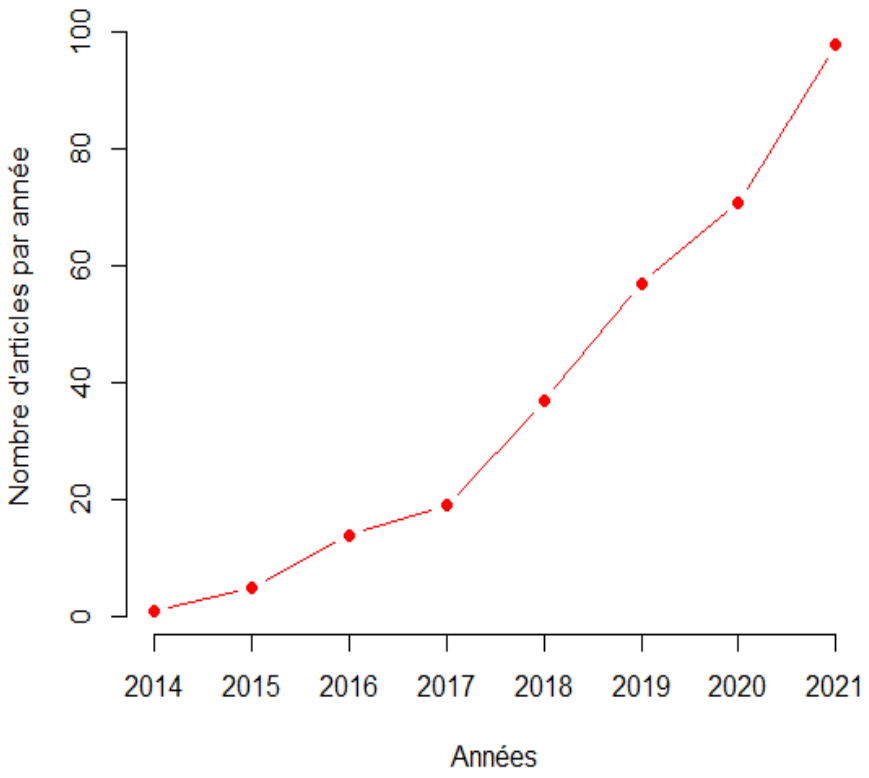

FIGURE 7 – Graphique représentant le nombre d'articles utilisant le terme *Ontolex Lemon* par année
Source: requêtes effectuée sur google scholar triées par année

Toutefois, le choix de ce modèle repose sur des critères qui peuvent peut être contestés. Le Specialist Lexicon est, dans une certaine mesure, un modèle pertinent pour la gestion lexicale. Il est conçu pour fonctionner de pair avec le Lexical Tools,

qui met à disposition de nombreux outils d'aide à la gestion lexicale.

6.1.1 Linguistic Linked Open Data cloud

Le Linguistic Linked Open Data cloud (LLOD) [47] est le fruit du travail d'un regroupement d'experts de différents domaines en liens avec la linguistique, l'Open Linguistics Working Group (OWLG) [48]. L'OWLG travaille autour des notions de l'ouverture des données des ressources linguistiques, de la centralisation des méthodes et travaux autour du domaine de la linguistique. Le LLOD est donc la collection de différentes entités portant sur la linguistique. On retrouve dans le LLOD des bases de données lexicales, comme WordNet [40] et DBpedia [41].

WordNet contient en autre des données lexicales, dont les entités sont regroupées au sein de *synset*, qui sont ensemble de termes synonymes. Initialement uniquement disponible pour la langue anglaise, il existe aujourd'hui des ressources dans d'autres langages, bien que moins riches, et notamment en français, comme le Wordnet Libre du Français (WOLF) [49].

DBpedia a pour but d'extraire et structurer les informations multi-langues de Wikipedia. De nombreuses relations sémantiques y sont contenues, ainsi que la traduction entre les langues des termes. DBpedia est un point central dans le Web des Données [47]. DBpedia contient des millions d'entités de domaines différents décrits à l'aide de triplets RDF.

Les données du LLOD sont reliées au vocabulaire Ontolex-Lemon, ce qui renforce la pertinence de ce modèle pour son utilisation dans la gestion des ressources lexicales.

6.2 Contrôle de la structure

Nous avons proposé un algorithme qui permet de garantir le respect des cardinalité en fonction des terme présent dans les données d'origine. Cet algorithme a été évalué avec succès sous la forme d'un test simple (tableau 2 et 3)

6.3 Services et méthodes

Il est intéressant de constater que les termes polysémiques proviennent majoritairement de la présence de plusieurs RTO dans la base de données graphes. Individuellement, ces RTO présentes peu de LexicalEntry polysémiques (18), ce nombre est multiplié par 10 losqu'on les regroupe (1 829). La majorité des termes polysémiques provient donc du fait que ces différentes RTO utilisent les mêmes termes.

Prendre la dimension conceptuelle dans la méthode de détection des termes polysémiques permettrait probablement de reconsidérer de nombreux termes catégorisés initialement comme polysémique en réalité comme non ambiguë. En effet, le concept C189+0 de la CIM-10, 144 d'OrphaNet et D003123 du MeSH sont sans doute lié par une relation skos:exactMatch, *Syndrome de Lynch* n'est sans doute pas à considérer comme un terme polysémique.

Concernant les termes polysémiques retrouvés dans la CIM-10, *Syphilis cardiovasculaire* est le terme utilisé pour représenter deux concepts : A520 et I980. Le

premier porte dans la classification un obèle (A52.0†), ce concept porte donc sur la maladie de manière générale, tandis ce que le second porte un astérisque (I98.0*), ce concept porte donc sur une manifestation clinique.

Malgré que *Syphilis cardio-vasculaire* soit considéré comme polysémique avec la méthode employée, la distinction des deux concepts n'a sans doute que peu d'importance pour les praticiens hospitaliers, une relation d'un point de vue conceptuelle parviendrait sans doute à recatégoriser ce terme de polysémique à non-ambiguë.

Pour les concepts R090 et T71, il est intéressant que les termes utilisés en anglais sont différents, contrairement au français. En effet, *Asphixia* et *Asphyxiation* sont respectivement utilisés en anglais, ou *Asphixie* l'est pour les deux en français. De la même manière que pour *Syphilis cardio-vasculaire*, ces concepts sont probablement assez proches pour que l'on considère que leurs distinctions ne soient pas pertinentes d'un point de vue du métier de praticien hospitalier.

6.4 Critique d'Ontolex-Lemon

Ontolex-Lemon a donc été publié dans une version finale. Son core est le fruit d'un travail de plusieurs années et semble être suffisamment abstrait pour qu'il puisse se coupler avec des modules complémentaires pour traiter les problèmes de gestion du domaine lexical. Bien que complet, ce modèle n'en reste pas moins complexe et des compétences dans le domaine de la linguistique sont nécessaires afin de bien cerner les différents enjeux et utilités de toutes ses parties. De plus, les modules complémentaires ne permettent pas aujourd'hui de résoudre certaines problématiques de la gestion lexicale. Par exemple, le module *decomp* est limité pour la description morphologique des LexicalEntry dans sa version aujourd'hui publiée[50].

Malgré les limites que peuvent présenter les modules aujourd'hui, il est envisageable qu'elles soient levées par l'introduction de nouveaux modules, tel que le module *Morphology* pour pallier les limites du module *decomp*, en permettant la description d'un lemme en morphème et ainsi gagner en précision grammairienne pouvant aider au développement de services utiles pour le TAL.

La figure 7 permet de visualiser l'augmentation du nombre d'articles publiés par année où le terme *Ontolex Lemon* apparaît. Cette augmentation laisse espérer que son utilisation se développe et que des outils proposant des méthodes abstraites de gestion lexicale voient le jour.

Il est tout de même à noter que des projets récents autour d'Ontolex-Lemon existent et se rapprochent des objectifs de ce projet. Aucun outil permettant de répondre exactement à notre besoin n'a été trouvé, mais il est pertinent de les citer afin de mieux cerner l'utilité d'Ontolex-Lemon.

Ainsi, des projets permettant la conversion de données au format TermBase eXchang (TBX), une norme de représentation des données conceptuelles des terminologies [51], vers un format RDF selon le modèle Ontolex-Lemon existent [52]. Cela ne répond pas directement à notre problématique car centré sur TBX, mais on peut constater que la transformation vers Ontolex-Lemon de différentes sources existe et

que ce modèle permet donc à priori de faire converger des standards hétérogènes pour le champ de la gestion lexicale.

Un projet open source permettant la création de données au format Ontolex-Lemon est également disponible : LexO [53]. À la différence de ce qui est présenté dans ce mémoire, LeXO ne permet pas une automatisation de la conversion de RTO vers le modèle Ontolex-Lemon, mais une approche plus fine et manuelle de la description lexicale de termes selon le modèle Ontolex-Lemon.

Enfin, la conversion du *Dictionnaire étymologique de l'ancien français*[4] (DEAF) a été réalisée [54]. Bien que manuelle, cette transformation appuie sur le fait qu'Ontolex-Lemon est adapté à des lexiques divers et non seulement médicaux.

4. http ://www.deaf-page.de/fr/

7 Conclusion

La mise en place d'une structure de gestion lexicale basé sur le modèle Ontolex-Lemon semble permettre, bien que ce modèle ne soit pas destiné à l'unique gestion du lexique des RTO du domaine biomédical, de développer des outils et méthodes d'aides aux processus de TAL pour ce domaine, et donc d'aider à la RI dans documents en texte libre présents trouvables dans des EDS tel que l'EDS du CHU de Bordeaux.

L'implémentation de ce modèle en une structure de gestion lexicale n'a pas résolu tous les défis posés par les difficultés de ce modèle, notamment ceux concernant la gestion des différentes Form des LexicalEntry, nécessitant des méthodes spécifiques pour le moment non développées. La suite du projet doit donc prendre en compte cette limite et s'atteler à y remédier.

La seule partie d'Ontolex-Lemon implémentée est son core. Le module *decomp* semble être implémentable et permettrait sans doute d'enrichir d'informations sur les potentiels synonymes, en développant des outils d'enrichissements lexicaux à partir de la manière de décomposer des LexicalEntry. L'enrichissement lexical permet d'améliorer les processus de RI [15]. Ce module permettrait également la détection de termes polysémiques, en détectant les lemmes décomposables en sous-lemmes qui seraient eux même ambiguës. La polysémie est un problème en NLP et impact la précision de ses résultats [14].

Certains outils du Lexical Tools [38], comme le STMT, semble fournir les méthodes permettant de décomposer les termes, il serait intéressant de se pencher sur une possible utilisation pour traiter les données afin de les implémenter dans le module *decomp*.

Les outils et méthodes développés durant ce projet ne se basant que sur les données de la partie lexicale des RTO, il est nécessaire d'avoir les données de la partie conceptuelle de ces même RTO afin de pouvoir développer des outils pertinents pour aider les processus de TAL. En couplant le modèle Ontolex-Lemon pour la partie lexicale des RTO avec le modèle SKOS pour la partie conceptuelle des RTO, il est probable qu'à l'aide de relations de type skos:exactMatch, skos:narrower ou skos:borader, il soit possible d'étoffer et d'améliorer les services implémentés actuellement. Les outils développés à l'aide de la partie conceptuelle doivent permettre de déterminer si un terme est réellement polysémique, ainsi que de distinguer selon le contexte qu'*IRC* est le signifié d'*insuffisance rénale chronique* ou d'*insuffisance respiratoire chronique* par exemple.

L'utilité de cette structure et des services qui y sont développés ne sera réellement évalué que lorsque des méthodes de TAL utilisant des outils implémentés dans cette structure seront utilisées et analysées. Pour que cette étape voit le jour, il est nécessaire au préalable d'améliorer cette structure à l'aide de nouveaux outils et de l'implémentation de nouveaux modules d'Ontolex-Lemon.

8 Références / Bibliographie

Références

[1] Willem G van Panhuis, Proma Paul, Claudia Emerson, John Grefenstette, Richard Wilder, Abraham J Herbst, David Heymann, and Donald S Burke. A systematic review of barriers to data sharing in public health. BMC Public Health, 14(1):1144, December 2014.

[2] R. Griffier, V. Jouhet, F. Thiessard, and S. Cossin. Identification des verrous et des leviers à la réutilisation secondaire des données dans un établissement de santé. Revue d'Épidémiologie et de Santé Publique, 68:S49–S50, March 2020.

[3] F. J. Martin-Sanchez, V. Aguiar-Pulido, G. H. Lopez-Campos, N. Peek, and L. Sacchi. Secondary Use and Analysis of Big Data Collected for Patient Care. Yearbook of Medical Informatics, 26(1):28–37, August 2017.

[4] A Hansell, Alex Bottle, L Shurlock, and Paul Aylin. Accessing and Using Hospital Activity Data. Journal of public health medicine, 23:51–6, March 2001.

[5] Gabriel A. Brat, Griffin M. Weber, Nils Gehlenborg, Paul Avillach, Nathan P. Palmer, and Luca. Chiovato et al. International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium. NPJ digital medicine, 3:109, 2020.

[6] Jeffrey G. Klann, Hossein Estiri, Griffin M. Weber, Bertrand Moal, Paul Avillach, and Chuan. Hong et al. Validation of an internationally derived patient severity phenotype to support COVID-19 analytics from electronic health record data. Journal of the American Medical Informatics Association: JAMIA, 28(7):1411–1420, July 2021.

[7] Griffin M. Weber, Harrison G. Zhang, Sehi L'Yi, Clara-Lea Bonzel, Chuan Hong, and Paul. Avillach et al. International Changes in COVID-19 Clinical Trajectories Across 315 Hospitals and 6 Countries: Retrospective Cohort Study. Journal of Medical Internet Research, 23(10):e31400, October 2021.

[8] Harrison G. Zhang, Arianna Dagliati, Zahra Shakeri Hossein Abad, Xin Xiong, Clara-Lea Bonzel, and Zongqi. Xia Scott. International electronic health record-derived post-acute sequelae profiles of COVID-19 patients. NPJ digital medicine, 5(1):81, June 2022.

[9] Chuan Hong, Harrison G. Zhang, Sehi L'Yi, Griffin Weber, Paul Avillach, and Bryce W. Q. . Tan et al. Changes in laboratory value improvement and mortality rates over the course of the pandemic: an international retrospective cohort study of hospitalised patients infected with SARS-CoV-2. BMJ open, 12(6):e057725, June 2022.

[10] Ralph Grishman. Information Extraction. IEEE Intelligent Systems, 30(5):8–15, September 2015. Conference Name: IEEE Intelligent Systems.

[11] Stefan Buttcher, Charles L. A. Clarke, and Gordon V. Cormack. Information Retrieval: Implementing and Evaluating Search Engines. MIT Press, February 2016. Google-Books-ID: 2c3RCwAAQBAJ.

[12] Aurélie Névéol. Traitement Automatique de la Langue Biomédicale. page 92.

[13] Priya H. Dedhia, Kallie Chen, Yiqiang Song, Eric LaRose, Joseph R. Imbus, Peggy L. Peissig, Eneida A. Mendonca, and David F. Schneider. Ambiguous and Incomplete: Natural Language Processing Reveals Problematic Reporting Styles in Thyroid Ultrasound Reports. Methods of Information in Medicine, January 2022. Publisher: Georg Thieme Verlag KG.

[14] Noriko Tomuro and Steve Lytinen. Polysemy in Lexical Semantics – Automatic Discovery of Polysemous Senses and Their Regularities. page 1.

[15] Souheyl Mallat, Anis Zouaghi, Emna Hkiri, and Mounir Zrigui. Method of Lexical Enrichment in Information Retrieval System in Arabic. International Journal of Information Retrieval Research (IJIRR), 3(4):35–51, October 2013. Publisher: IGI Global.

[16] Vitor D. T. Andrade, Pedro Ruas, and Francisco M. Couto. Named Entity Recognition and Linking: a Portuguese and Spanish Oncological Parallel Corpus, September 2021. Pages: 2021.09.16.460605 Section: New Results.

[17] Maud Ehrmann. Les Entitées Nommées, de la linguistique au TAL: Statut théorique et méthodes de désambiguïsation. page 296.

[18] Yannick Prié and Serge Garlatti. Méta-données et annotations dans le Web sémantique. January 2004.

[19] Eneko Agirre, Oier Lopez de Lacalle, Aitor Soroa, and Informatika Fakultatea. Knowledge-Based WSD on Specific Domains: Performing Better than Generic Supervised WSD. page 6.

[20] Zhao-Yan Ming and Tat Seng Chua. Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling. Information Sciences, 307:18–38, June 2015.

[21] Shaidah Jusoh. A STUDY ON NLP APPLICATIONS AND AMBIGUITY PROBLEMS. . Vol., (6):14, 2005.

[22] Pierre Zweigenbaum. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. page 23.

[23] Didier Bourigault and Nathalie Aussenac-Gilles. Construction dontologies à partir de textes. page 22.

[24] SKOS eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant, 18 August 2009 Recommendation Edition.

[25] About: Lexikon. https://dbpedia.org/page/lexicon.

[26] N. F. de Keizer, A. Abu-Hanna, and J. H. Zwetsloot-Schonk. Understanding terminological systems. I: Terminology and typology. Methods of Information in Medicine, 39(1):16–21, March 2000.

[27] Louise Guthrie, James Pustejovsky, Yorick Wilks, and Brian M. Slator. The role of lexicons in natural language processing. Communications of the ACM, 39(1):63–72, January 1996.

[28] Kavi Mahesh and Sergei Nirenburg. Knowledge-based systems for natural language processing. January 1997.

[29] Stan C Kwasny and Barry L Kalman. The Case of the Unknown Word:. page 5.

[30] P. Cimiano, P. Buitelaar, J. McCrae, and M. Sintek. LexInfo: A declarative model for the lexicon-ontology interface. Journal of Web Semantics, 9(1):29–51, March 2011.

[31] John McCrae, Guadalupe Cea, Paul Buitelaar, Philipp Cimiano, Thierry Declerck, Asuncion Gomez-Perez, Jorge Gracia, Laura Hollink, Elena Montiel-Ponsoda, Dennis Spohr, and Tobias Wunner. Interchanging lexical resources on the Semantic Web. Language Resources and Evaluation, 46:701–719, December 2012.

[32] Domenico M. Pisanelli, Aldo Gangemi, Massimo Battaglia, and Carola Catenacci. Coping with Medical Polysemy in the Semantic Web: the Role of Ontologies. MEDINFO 2004, pages 416–419, 2004. Publisher: IOS Press.

[33] Eclipse RDF4J developers. Welcome · Eclipse RDF4J™ | The Eclipse Foundation. https://rdf4j.org/.

[34] JHipster - Full Stack Platform for the Modern Developer! https://www.jhipster.tech/.

[35] Emilio Rubiera, Luis Polo, Diego Berrueta, and Adil El Ghali. TELIX: An RDF-Based Model for Linguistic Annotation. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Elena Simperl, Philipp Cimiano, Axel Polleres, Oscar Corcho, and Valentina Presutti, editors, The Semantic Web: Research and Applications, volume 7295, pages 195–209. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. Series Title: Lecture Notes in Computer Science.

[36] Amanda Payne and Destinee Tormey. Lister Hill National Center for Biomedical Communications National Library of Medicine Bethesda, Maryland. page 89.

[37] Olivier Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(Database issue):D267–270, January 2004.

[38] Chris J Lu, Amanda Payne, and James G Mork. The Unified Medical Language System SPECIALIST Lexicon and Lexical Tools: Development and applications. Journal of the American Medical Informatics Association, 27(10):1600–1605, October 2020.

[39] Jeff Williamson. Development of Sub-Term Mapping Tools (STMT). page 1.

[40] Mariano Sigman and Guillermo A. Cecchi. Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences of the United States of America, 99(3):1742–1747, February 2002.

[41] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web, 6(2):167–195, 2015.

[42] The lemon cookbook. https://lemon-model.net/lemon-cookbook/index.html.

[43] John McCrae, Elena Montiel-Ponsoda, and Philipp Cimiano. Integrating WordNet and Wiktionary with lemon. Linked Data in Linguistics, pages 25–34, 2012.

[44] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, and Jeff Pan, editors, The Semantic Web: Research and Applications, volume 6643, pages 245–259. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. Series Title: Lecture Notes in Computer Science.

[45] John P McCrae, Julia Bosque-Gil, Jorge Gracia, and Paul Buitelaar. The OntoLex-Lemon Model: Development and Applications. page 11.

[46] John McCrae, Dennis Spohr, and Philipp Cimiano. Linking Lexical Resources and Ontologies on the Semantic Web with Lemon. volume 6643, pages 245–259, July 2011.

[47] Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. Towards a Linguistic Linked Open Data cloud : The Open Linguistics Working Group. TAL, 52:245–275, January 2011.

[48] Open linguistics. https://linguistics.okfn.org/index.html.

[49] Benoît Sagot and Darja Fišer. Construction d'un wordnet libre du français à partir de ressources multilingues. page 11.

[50] Bettina Klimek, John P McCrae, Julia Bosque-Gil, Maxim Ionov, James K Tauber, and Christian Chiarcos. Challenges for the Representation of Morphology in Ontology Lexicons. page 22, 2019.

[51] Félix Castro, Alexander Gelbukh, Miguel González, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, and Gerhard Weikum, editors. Advances in Artificial Intelligence and Its Applications: 12th Mexican International Conference on Artificial Intelligence, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part I, volume 8265 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[52] Silvia Piccini and Federica Vezzani. Entre TBX et Ontolex-Lemon : Quelles Nouvelles Perspectives en Terminologie ? page 4.

[53] Andrea Bellandi. LexO: an open-source system for managing OntoLex-Lemon resources. Language Resources and Evaluation, 55(4):1093–1126, December 2021.

[54] Sabine Tittel and Christian Chiarcos. Historical Lexicography of Old French and Linked Open Data: Transforming the resources of the Dictionnaire étymologique de l'ancien français with OntoLex-Lemon. May 2018.

[55] 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 edition.

[56] Maurice Grevisse. Grammaire française. Duculot-gembloux et hatier-paris edition, 1964.

[57] Erkki Luuk. Syntax–Semantics Interface. International Encyclopedia of the Social & Behavioral Sciences, December 2015.

[58] Christian Touratier. Chapitre III. La morphologie. In Morphologie et morphématique : Analyse en morphèmes, Langues et langage, pages 61–69. Presses universitaires de Provence, Aix-en-Provence, December 2012. Code: Morphologie et morphématique : Analyse en morphèmes.

[59] Fiammetta Namer, Robert Baud, Anita Burgun, Stéfan J. Darmoni, Natalia Grabar, Eric Jarrousse, Franck Le Duff, Patrick Ruch, Benoît Thirion, and Pierre Zweigenbaum. UMLF : construction d'un lexique médical francophone unifié. September 2003.

[60] Hamada A Nayel, H L Shashirekha, Hiroyuki Shindo, and Yuji Matsumoto. Improving Multi-Word Entity Recognition for Biomedical Texts. page 13.

[61] Dictionnaire de l'Académie Nationale de Médecine.

[62] Shamim Ara Mollah and Stephen B. Johnson. Automatic learning of the morphology of medical language using information compression. AMIA ... Annual Symposium proceedings. AMIA Symposium, page 938, 2003.

9 Annexes

9.1 Définitions des notions abordées par le projet

Afin de comprendre au mieux l'objectif de ce projet, une liste de différentes notions abordées est présentée ci après :

Concept et termes : les notions de concepts et de termes sont difficiles à appréhender individuellement, elles sont définies grâce au triangle sémiotique d'Ogden et Richards [55]:

Un concept est une idée, une notion référençant un objet. Ainsi, les concepts sont les idées exprimées par des termes, et les termes permettent d'exprimer les idées représentées par les concepts. Les termes n'ont de sens que par les concepts auxquels ils réfèrent, et l'on parle de signifiés (concepts) et de signifiants (termes). Un terme sans concept n'aurait pas de sens, et un concept sans termes ne pourrait être exprimé.

Grammaire : elle correspond à l'étude d'une langue [56]. La grammaire se décompose en deux disciplines complémentaires que sont l'étude de la syntaxe et de la morphologie.

Syntaxe : elle étudie les relations entre les mots et la manière dont les mots s'assemblent dans une phrase [57]. C'est la syntaxe qui dicte entre autre qu'en français un sujet se place avant le verbe qui se trouve lui-même devant le complément d'objet direct. Ainsi, *le médecin a prescrit un antibiotique* est correcte syntaxiquement, *a prescrit le médecin un antibiotique* ne l'est pas. La syntaxe décrit également que *le médecin* est sujet de *a prescrit* qui a pour complément d'objet direct *un antibiotique*.

Morphologie : elle étudie les mots [58]. La morphologie décrit la composition d'un mot, ainsi que son sens, sa signification (le concept auquel il est rattaché).

Lexique : il correspond à l'ensemble des mots d'une langue ou d'un domaine de connaissances [59]. En linguistique, les mots composants un lexique sont nommés des lemmes. La notion de domaine de connaissance est nécessaire car si le dictionnaire de l'académie française de médecine est un lexique, il y est trouvable des lemmes que le dictionnaire de la langue française ne décrit pas. Les dictionnaires médicaux contiennent de nombreux termes médicaux qui sont dits composés [60].

Ainsi, il est trouvable dans le dictionnaire de l'académie de médecine français des lemmes tel qu *insuffisance cardiaque après remplissage vasculaire* [61]. Bien que non présent dans le dictionnaire de la langue française, il est considéré comme un lemme en regard du lexique utilisé. En revanche, ce lemme est composé de différents sous-lemmes retrouvés dans le dictionnaire de la langue française : *insuffisance cardiaque* (ou *insuffisance* et *cardiaque*), *remplissage* et *vasculaire*.

Sémantique : elle correspond au sens, à la signification d'une entité linguistique [57]. La morphologie des mots et la syntaxe influence la sémantique d'une phrase. Il est nécessaire de savoir comment les mots interagissent entre eux et les concepts auxquels ils renvoient pour comprendre le contexte d'un document. La syntaxe et la morphologie donne donc la sémantique. La phrase *Paul appelle Jean* a un sens, une sémantique, donnée par la morphologie qui fournit la signification de chaque mot présent dans la phrase. La phrase *Jean appelle Paul* possède une sémantique différente, changé par la syntaxe, puisque l'ordre des mots est différent. Pour comprendre une phrase, lui attribuer un contexte, la morphologie et la syntaxe est nécessaire.

Composition d'un lemme : L'unité élémentaire (et donc indivisible) d'un lemme d'un point de vue morphologique se nomme un morphème [62].

Un morphème présent dans le lexique est nommé un lexème. Prenons le lexique des termes médicaux en français. Dans ce lexique se trouve comme entrée lexicale *génital*. Cette entrée lexicale est un morphème puisqu'on ne peut pas le décomposer morphologiquement, c'est également un lexème puisque c'est un morphème présent dans le lexique. Les lexèmes sont de fait des lemmes, puisque ce sont des entrées lexicales.

Le morphème *con* de *congénital* est dit morphème dérivationnel. C'est un morphème non présent dans le lexique des termes médicaux en français, il n'existe pas seul, mais est capable associé à des lexèmes de changer le concept auquel le lemme réfère.

Un morphème flexionnel est un morphème qui ajouté à un lemme ne change pas le concept auquel le lemme réfère. Ainsi, *congénitale* est une flexion du lemme *congénital* car il est constitué du morphème flexionnel *e*, sans changer le concept auquel *congénital* se réfère. il est également dit que *congénital* est la fore canonique du lemme.

Un terme peut être une forme canonique ou fléchie d'un lemme.

Un lemme peut être constitué de plusieurs sous-lemmes, comme *atrésie congénitale des voies biliaires*, qui peut être décomposé de différentes façon en fonction de la présence ou non des sous-lemmes dans le lexique.

9.2 Synonymie, enrichissement lexical et ambiguïté

Synonymie : Deux lemmes sont définis comme étant synonymes s'ils partagent pour signifié un même concept. Par exemple, les lemmes *Gastralgie* et *Douleurs à l'estomac* sont synonymes, puisqu'ils sont les signifiants d'un même concept.

Enrichissement lexical : ce processus consiste à générer de nouveaux lemmes synonymes à un lemme.

Formes fléchies d'un lemme : le processus de flexion de lemmes en formes non canoniques ne fait pas à proprement parti de l'enrichissement lexical, puisque qu'une

entrée de lexique est par définition un lemme, et non pas une forme fléchie d'un lemme.

Ambiguïté : un lemme, ou l'une de ses formes fléchies, est dit ambiguë s'il est le signifiant de plusieurs concepts. Par exemple, le lemme *IVG* est ambiguë car il possède plusieurs signifiés: il peut avoir un sens cardiologique, avec le concept d'insuffisance ventriculaire gauche, ou gynécologique, avec le concept d'interruption volontaire de grossesse. Le terme polysémie est également employé pour parler d'ambiguïté.

Résumé

L'utilisation de services se basant sur un modèle de gestion lexicale permettrait d'apporter une aide aux processus de recherche d'informations dans les documents en texte libre. Ce document est la synthèse d'un travail réalisé au sein de l'unité IAM du CHU de Bordeaux, ayant pour but l'implémentation d'une structure se basant sur un modèle de gestion lexicale afin que des services et méthodes puissent y être développés. Le choix du modèle de gestion lexicale s'est porté sur Ontolex-Lemon, qui répondait le mieux à nos attentes. L'évaluation de cette structure a été faite en y intégrant quatre ressources termino-ontologiques, et des services et méthodes comme la détection de termes synonymes et polysémiques y ont été intégrés.

Mots-clefs : Science de l'informatique médicale, Sciences de l'information, Traitement automatique du langage naturel, Entrepôt de données.

Abstract

The use of services based on a lexical management model would help in the process of searching for information in free text documents. This document is the synthesis of a work carried out within the IAM unit of the CHU of Bordeaux, with the aim of implementing a structure based on a lexical management model so that services and methods can be developed there. The choice of the lexical management model was Ontolex-Lemon, which best met our expectations. The evaluation of this structure was done by integrating four termino-ontological resources, and services and methods such as synonym and polysemous terms detection were integrated.

Keywords : Medical Informatics, Information Science, Natural Language Processings, Data Warehousing.