Université de Rouen U.F.R. des Sciences et Techniques
École doctorale Sciences Physiques, Mathématiques et de l'Information pour l'Ingénieur

THÈSE

Pour obtenir le grade de
Docteur en informatique de l'université de Rouen

Modélisation, réalisation et évaluation d'un portail Multi-terminologique, Multi-discipline, Multi-lingue (3M) dans le cadre de la Plateforme d'Indexation Régionale (PlaIR)

Julien GROSJEAN

Soutenue le 17 octobre 2014

Soutenue devant le jury composé de :

Directeur de thèse :	Pr Stéfan Darmoni	PU-PH	CHU de Rouen
Co-encadrante :	Dr Lina Soualmia	MCF	Université de Rouen
Rapportrices :	Dr Catherine Duclos	MCU-PH (HDR)	AP-HP & Université Paris 13
	Pr Chantal Reynaud	PU	Université Paris 11
Président du jury :	Pr Thierry Paquet	PU	Université de Rouen
Examinatrice :	Dr Marie-Christine Jaulent	DR	Université Paris 6 et 13

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes

Résumé

Les Systèmes d'Organisation des Connaissances (SOC) sont des vocabulaires, plus ou moins structurés qui sont surtout utilisés dans les Systèmes d'Information. Qu'il s'agisse d'activités de recherche, d'indexation de documents, de codage de dossiers informatisés patient ou encore d'enseignement, ces SOC proposent des usages très différents et spécifiques de domaines. Principalement utilisés en sciences, ils sont aujourd'hui des ressources permettant de stocker la connaissance mais aussi de l'exploiter.

La possibilité de gérer simultanément plusieurs SOC de domaines offre de nombreux avantages. Leurs contenus sont très riches et, unis, il permettent de créer des réseaux sémantiques ou des super-ensembles de connaissance. Le travail présenté ici s'inscrit dans le cadre du projet PlaIR (Plateforme d'Indexation Régionale) qui vise notamment à construire un serveur Multi-terminologique, Multi-discipline et Multi-lingue (3M) pour aider les documentalistes, professionnels de santé, chercheurs, étudiants et experts à chercher et utiliser des SOC. Pour cela, notre approche a consisté à créer un méta-modèle 3M capable d'intégrer n'importe quel SOC puis de créer le serveur correspondant. Ce serveur (S3M) a été implémenté via la conception d'un schéma de base de données générique puissant et souple permettant d'y intégrer des informations hétérogènes de domaines différents. Il a ensuite été question d'inclure un certain nombre de SOC dans ce S3M à partir de choix et besoins spécifiques pour la communauté. Le principe est de fournir un service accessible sur l'Internet, pour les humains et les machines, avec un contenu enrichi. Pour cela, un site web, baptisé HeTOP (pour Health Terminology/Ontology Portal), a été développé et offre des accès sécurisés et personnalisés aux SOC dans un environnement adapté à une utilisation quotidienne intensive. De plus, afin de répondre au mieux aux besoins de la communauté, un Service Web dispense les mêmes fonctionnalités que HeTOP pour un accès par des machines. Enfin, il a été question de versionnage et d'édition de SOC du S3M via la conception de méthodes et d'outils capables non seulement de gérer les modèles et concepts des SOC mais aussi l'aide à la saisie pour les experts et la gestion de flux terminologiques.

Mots clés : terminologie, ontologie, modélisation, vocabulaire contrôlé, portail terminologique, portail ontologique, base de données.

Abstract

The Knowledge Organization Systems (KOS) are defined as vocabularies, more or less structured and are mainly used in Information Systems. These KOS are used for research, documents indexing, electronic health records coding or teaching. They are offering various different usages in specific domains. Mainly used in science, they are now resources for storing and processing knowledge.
Managing simultaneously multiple KOS of close domains presents many advantages. Their contents are rich and united, it can create semantic networks or super-sets of knowledge resources. The hereby work is a part of the PlaIR (Plateforme d'Indexation Régionale) project which aims to build a Multi-terminological Multi-discipline and Multi-lingual (3M) server to help librarians, health professionals, researchers, students and experts to search, browser and use KOS. Therefore, our approach consisted in creating a 3M meta-model able to integrate any KOS and then creating the corresponding server. This server (S3M) was implemented by designing a scheme of a generic database which is powerful and flexible and allows to integrate heterogeneous data from different domains. The next step consisted in integrating various KOS in this S3M which were selected using several specific criteria. The final goal is to provide an accessible service on the Internet for both humans and machines. To do so, a website called HeTOP (for Health Terminology/Ontology Portal) was developped and offers now a secured and personalized access to KOS in a suitable environment for an intensive daily use. In order to best meet the needs of the community, a Web Service provides the same functionalities as HeTOP for machines. Finally, other problematics have been studied such as KOS versionning and edition via the design of methods and tools allowing to manage both KOS models and concepts but also via input help systems for experts and management of terminological flows.

Keywords : terminology, ontology, modelling, controlled vocabulary, terminological portal, ontological portal, database.

Remerciements

En premier lieu, je souhaite vivement remercier le Professeur Stéfan Darmoni, toujours à l'écoute, motivant et véritable moteur de l'équipe et de ses travaux. Je le remercie également pour m'avoir donné toute sa confiance et ce, dès notre première rencontre. Je tiens ensuite à remercier Lina Soualmia pour son encadrement et ses conseils avisés.

Je remercie également le Docteur Catherine Duclos et le Professeur Chantal Reynaud qui ont accepté d'évaluer mon travail.

Un grand merci à toute l'équipe CISMeF, pour le soutien et la bonne humeur quotidienne. En particulier, j'adresse un grand merci à mes compères du midi : Gaétan « Geeko » Kerdelhué, Cédric « Dota2 » Queinnec, Nicolas « Plomb V » Griffon et mes deux compagnons de bureau qui m'ont supporté (et me supportent encore) : Badisse « WKZ » Dahamna et Ivan « Moizez » Kergourlay. Merci également à tout le reste de l'équipe CISMeF (et anciens ou associés) pour leur sympathie : Tayeb, Adila, Catherine, Benoit, Mher, Sandrine, Josette, Chloé, Romain, Ahmed Diouf et les Dr Jean-Philippe Leroy et Philippe Massari. Sans oublier les étudiants qui ont également contribué à ce travail : Guillaume, Nelson, Antoine, Philippe, Kevin, Victor, Quentin et Teng.

Enfin, je souhaite remercier ma famille, particulièrement Estelle et Antoine qui me subissent (depuis bien longtemps) et mes parents qui m'ont insufflé curiosité et ténacité, éléments indispensables à la pratique de la Recherche.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

API Application Programming Interface

BT Broader Term : terme plus large

CCAM Classification Commune des Actes Médicaux : référentiel français

CISMeF Catalogue et Index des Sites Médicaux de langue Française

CUI Concept Unique Identifier : identifiant unique de l'UMLS

DBGUI DataBase Graphic User Interface : outil de saisie du SI CISMeF

DP Datatype Property : attribut d'objet dans la base de données générique

DPI Dossier Patient Informatisé

HeTOP Health Terminology/Ontology Portal

HL7 Health Level 7 : organisation qui définit des normes

MDP Model Datatype Property : attribut de TYPE_ID dans la base de données générique

MeSH Medical Subject Headings : thésaurus documentaire en médecine

MIN Model Inheritance : relation d'héritage entre TYPE_ID dans la base de données générique

MOP Model Object Property : relation entre TYPE_ID dans la base de données générique

NT Narrower Term : terme plus étroit

OP Object Property : relation entre objets dans la base de données générique

PlaIR Plateforme d'Indexation Régionale

PT Prefered Term : terme préféré

RDF Resource Description Framework : modèle de description de données

S3M Serveur Multi-terminologique Multi-discipline Multi-lingue

SGBD Système de Gestion de Base de Données

SOC Système d'Organisation des Connaissances

SQL Structured Query Language : pour requêter des SGBD

SW Service Web : programme pour échanger des données sur le web

TI TYPE_ID : type élémentaire dans la base de données générique

UMLS Unified Medical Language System

Chapitre 1

Introduction

Sommaire

1.1 Organiser les connaissances

Représenter, stocker ou structurer les connaissances sont des problématiques très anciennes. Les dessins, premières traces de connaissance ont fait place aux écrits, plus ou moins structurés selon leurs buts. La pensée et la connaissance du monde étaient déjà répertoriées à l'Antiquité ; Pline l'Ancien avait d'ailleurs écrit une sorte d'encyclopédie afin de ne pas perdre la connaissance si difficile à accumuler et à transmettre (*Histoire naturelle* sur 37 volumes, entre 50 et 75 après J.C.). Au début de l'an 1000, Ibn Sînâ (Avicenne) rédige une œuvre monumentale sur la médecine dans son *Canon de la médecine* resté célèbre des siècles durant [Yahia, 1952]. Depuis, la façon de ranger et d'accéder à ces ressources n'a cessé d'évoluer. Les croquis détaillés et annotés ont été un support extraordinairement efficace pendant des centaines d'années. Entre temps, au niveau du langage et de son utilisation, l'encyclopédie et le dictionnaire ont été largement diffusés grâce à l'imprimerie. Les notions

d'ordre, de classification et même d'indexation sont apparues rapidement.
Aujourd'hui, les connaissances accumulées sont gigantesques. De plus, depuis l'avènement de l'Internet, leur accès est largement facilité et beaucoup de nouveaux problèmes se posent quant à leur organisation, leur stockage voire même leur véracité.

1.2 Contexte de recherche, projets InterSTIS et PlaIR

Les travaux présentés ici traitent de la gestion des vocabulaires contrôlés servant non seulement à stocker et organiser la connaissance mais aussi à l'exploiter. Dans ce contexte, plusieurs projets de recherche ont été menés et sont encore en cours. Plus particulièrement, ces problématiques d'ingénierie des connaissances sont très étudiées en informatique médicale. C'est d'ailleurs dans ce contexte particulier que s'inscrit cette thèse même si l'approche présentée tend à être plus générique. Je détaillerai ici deux projets dans lesquels cette étude a pris sa source puis je ferai une brève introduction aux deux disciplines prépondérantes à ces travaux.

1.2.1 Projet InterSTIS

InterSTIS (Interopérabilité Sémantique des Terminologies dans les Systèmes d'Information de Santé français) [1] est un projet de recherche financé par l'Agence Nationale de la Recherche (ANR-07-TecSan-10) [2] qui s'est déroulé entre 2008 et 2011 [Joubert *et al.*, 2012]. Coordonné par l'entreprise Vidal [3] et le LERTIM (Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale, Université de la Méditerranée à Marseille), ce projet avait pour objectif principal de centraliser des terminologies francophones en santé dans un serveur unique et de les rendre interopérables via la génération automatique d'alignements.
InterSTIS fut un succès, notamment au regard du démonstrateur développé pour rechercher et consulter les différentes terminologies intégrées. Cependant, même si le méta-modèle terminologique et les méthodologies d'intégration ont fait leurs preuves, les technologies utilisées ont montré des limites quant à une éventuelle industrialisation. Par exemple, le thésaurus MeSH (*Medical Subject Headings*, cf. 3.5.2) n'a pu être intégré à cause de son importante volumétrie. De plus, les temps de réponses de l'outil n'étaient pas adaptés à une utilisation quotidienne.
Fortement impliquée dans InterSTIS, l'équipe CISMeF Catalogue et Index des Sites Médicaux de langue Française, cf. 1.3.1) s'est appuyée sur les méthodologies développées dans ce projet de recherche pour créer son premier serveur multi-terminologique bilingue français/anglais. Les travaux présentés dans ce mémoire s'inscrivent natu-

1. http://cybertim.timone.univ-mrs.fr/recherche/projets-recherche/INTERSTIS/
2. http://www.agence-nationale-recherche.fr/
3. http://www.vidal.fr/

rellement dans la continuité d'InterSTIS mais également dans celle des travaux de la thèse de Pierre-Yves Vandenbussche intitulée *Définition d'un cadre formel de représentation des Systèmes d'Organisation de la Connaissance* [Vandenbussche, 2011] et financée par la société Mondeca [4], spécialiste dans la gestion et l'édition de terminologies.

1.2.2 Projet PlaIR

PlaIR (Plateforme d'Indexation Régionale) [5] est un projet de recherche co-financé par l'Union Européenne et la région Haute-Normandie (FEDER pour Fonds Européens de DÉveloppement Régional). Coordonné par l'équipe « DocApp » (Document et Apprentissage) du LITIS, PlaIR s'est déroulé de 2009 à 2012 et avait pour buts de mutualiser un ensemble de ressources documentaires numériques et numérisées et les bibliothèques logicielles d'analyse automatique ou semi-automatique de ces ressources pour constituer une plateforme d'indexation et de recherche d'information multi-domaines et multi-usages. Le premier axe du projet consistait en la reconnaissance de caractères imprimés sur des journaux anciens numérisés et en l'élaboration d'une plateforme de visualisation de ces documents. Le second axe avait pour but de réaliser une plateforme pour aider à l'indexation en multi-terminologie et en multi-discipline. Il s'agissait donc d'étendre le modèle multi-terminologique conçu lors du projet InterSTIS et consolidé par l'équipe CISMeF pour assurer ses fonctionnalités et sa validité à d'autres domaines que celui de la Santé et plus précisément pour le Droit du Transport et les Sciences de l'ingénieur.
Ces travaux de thèse s'inscrivent pleinement dans ce projet de recherche PlaIR en traitant par ailleurs d'autres problématiques connexes.

1.2.3 L'informatique médicale

Un système de santé a pour objectifs de mieux soigner les malades et de mieux gérer la distribution des soins [Ducrot & Dusserre, 1990]. Avec l'évolution des connaissances, des techniques et le perfectionnement des instruments, les informations dans le domaine de la médecine se complexifient. La pratique d'une médecine moderne et de qualité ne peut être dissociée d'un traitement rationnel de l'information médicale. L'informatique médicale est une science qui aide à recueillir les faits, à les mémoriser, à les échanger et à les interpréter [Degoulet & Fieschi, 1998]. Les Systèmes d'Organisation de la Connaissance occupent à ce titre une place privilégiée. Parmi les catégories d'utilisation, nous pouvons retenir :

– le codage de Dossiers Patient Informatisés (DPI) ;

4. http://www.mondeca.com/
5. http://www.plair.org/

- la normalisation d'un vocabulaire commun pour faciliter l'échange d'information ;
- l'inscription d'un volume de données qu'un cerveau ne peut mémoriser pour en extraire des tendances (*data mining* ou fouille de données) ;
- la formalisation des connaissances pour effectuer des tâches automatiques plus « intelligentes » par un ordinateur (extension de recherche par inférence).

1.2.4 L'ingénierie des connaissances

L'ingénierie des connaissances est une discipline issue de l'Intelligence Artificielle. Edward Feigenbaum et Pamela McCorduck la définissent en 1983 comme « la discipline consistant à intégrer de la connaissance dans les systèmes informatiques afin de résoudre des problèmes complexes nécessitant un haut niveau d'expertise humaine » [Feigenbaum & McCorduck, 1983]. Cette discipline regroupe plusieurs types d'activités comme la conception et le développement de bases de connaissances, l'intégration, la maintenance et l'évaluation des systèmes d'information, etc. L'apparition du web a considérablement élargi le champ d'action du domaine de l'ingénierie des connaissances. En effet, le partage et l'interopérabilité de la connaissance offrent de nombreux enjeux et problématiques tant sur les plans méthodologique que technique ; on peut notamment faire référence au Web de données [Bizer *et al.*, 2009] ou encore aux « Big Data » (données massives) [Manyika *et al.*, 2011].
Les Systèmes d'Organisation des Connaissances sont des produits de cette discipline et constituent un challenge important dans l'élaboration et l'exploitation des systèmes d'information documentaires.

1.3 Contexte de travail et projet de recherche

Dans cette section, je présente l'équipe CISMeF, ses travaux de recherche ainsi que les problématiques et objectifs de la présente étude.

1.3.1 L'équipe CISMeF

CISMeF est l'acronyme de Catalogue et Index des Sites Médicaux de langue Française (http://www.cismef.org/). Il s'agit initialement d'un projet visant à proposer des documents de qualité en santé pour les professionnels mais également pour les particuliers. Créée en 1995 par le Professeur Stéfan J. Darmoni et Benoit Thirion, cette structure fait partie du Centre Hospitalo-Universitaire de Rouen puis plus tard, du LITIS (Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes) dans l'équipe TIBS (Traitement de l'Information en Biologie Santé) faisant partie de l'Université de Rouen (http://www.litislab.eu/). En effet, en plus de son activité d'indexation et de mise à disposition des documents sur le web

(outil Doc'CISMeF, http://doccismef.chu-rouen.fr/), le CISMeF s'implique désormais dans des projets de recherche autour des problématiques du web sémantique, de l'e-santé et de l'informatique médicale.
Aujourd'hui constitué de 3 documentalistes dont une spécialiste en pharmacie, de 4 médecins, de 4 ingénieurs en informatique et d'un maître de conférence, le CISMeF est impliqué dans plusieurs projets ANR (Agence Nationale pour la Recherche) et dans nombre de projets annexes, liés à la Région Haute-Normandie, à l'Université de Rouen, ou même à des entreprises privées comme Vidal ou GlaxoSmithKline. La recherche au sein de l'équipe fait donc intervenir régulièrement des doctorants, post-doctorants et des étudiants ingénieurs d'écoles ou d'universités.

1.3.2 Le paradigme mono-terminologique

De nombreux systèmes documentaires reposent sur l'utilisation de terminologies ou d'ontologies. Parmi eux, CISMeF est un catalogue riche et de qualité contenant plus de 100 000 ressources de santé accessibles sur l'Internet. Ces documents sont indexés par des documentalistes experts dans le domaine de la santé et s'aident pour cela de plusieurs terminologies de référence comme le thésaurus MeSH (édité par la *National Library of Medicine* (NLM) aux États-Unis). Nous verrons par la suite que le MeSH est une ressource particulière en de nombreux aspects mais il s'agit surtout d'un vocabulaire internationalement reconnu pour indexer des documents scientifiques du domaine bio-médical ; le corpus de référence MedLine[6] (servi par son moteur de recherche PubMed) est d'ailleurs indexé grâce au MeSH.
De la même manière que MedLine, CISMeF s'est longtemps cantonné à l'utilisation d'une seule terminologie (entre 1995 et 2005). Cependant, comme l'a notamment montré Sakji [Sakji, 2010], pour des corpus scientifiques aussi vastes et pointus, le MeSH s'avère être trop restrictif et de nombreux concepts plus précis n'y sont pas définis. Ainsi, la décision de passer dans un univers multi-terminologique amène de nouvelles problématiques, tant méthodologiques que techniques.

1.3.3 Le passage en multi-terminologie

Alors que la NLM a décidé également d'étendre son vocabulaire à l'aide des « Concepts Supplémentaires » (remplaçant les « Concepts Chimiques Supplémentaires » introduits dans les années 1980) puis créé l'UMLS (Unified Medical Language System), CISMeF développe au début des années 2000 non seulement une extension du MeSH (appelée terminologie CISMeF) mais envisage également l'intégration dans son Système d'Information (SI) de nouvelles terminologies de référence, usuellement exploitées dans d'autres contextes. Parmi elles, la CIM-10, la SNOMED int. (3.5) pour les maladies et symptômes ou encore, l'ATC pour les médicaments.

6. *Medical Literature Analysis and Retrieval System Online*

L'utilisation simultanée de plusieurs terminologies dans un même système amène un grand nombre de problématiques liées à l'indexation, à la Recherche d'Information (RI), à son stockage et à son utilisation. En effet, comme démontré plus loin dans ce mémoire, les classifications ne sont pas toutes conçues pour des usages génériques mais correspondent plutôt à des utilisations spécifiques, dans des contextes bien précis. La multi-terminologie constitue en fait en soi un non-sens théorique ; elle est utilisée à des vues fonctionnelles et il faut bien comprendre ses inconvénients, ses dangers et ses enjeux pour en limiter les éventuels problèmes.

1.4 Objectifs

Mes objectifs ont donc été multiples autour d'un modèle Multi-terminologique Multi-discipline et Multi-lingue (abrégé en modèle 3M). Il s'agissait donc de modifier et de consolider le modèle initial (InterSTIS), de l'adapter à d'autres domaines (PlaIR) tout en intégrant un grand nombre de terminologies mais aussi d'ontologies au sein du système. En outre, un but important était de concevoir et développer l'application web exploitant toutes ces données terminologiques, tout en s'inscrivant dans une démarche de production à grande échelle, afin de proposer un service de qualité pour les centaines d'utilisateurs quotidiens de l'outil mono-terminologique existant auparavant. De plus, dans un monde scientifique de plus en plus diversifié mais pointu, un de mes objectifs était d'offrir aux utilisateurs du portail un service pertinent scientifiquement, multi-lingue (et inter-lingue), donc compréhensible par le plus grand nombre. En effet, même si l'anglais demeure aujourd'hui la langue pivot au niveau international, nombre de personnes, y compris des scientifiques, ne manipulent pas aussi bien l'anglais que leur langue maternelle, surtout pour des termes complexes appartenant à des champs disciplinaires bien précis.
En outre, il a fallu concevoir et développer des outils permettant l'édition des ressources terminologiques, à des fins de traductions, d'alignements, etc.
Par ailleurs, une réflexion sur la gestion des versions et des historiques des terminologies et de leurs données a été lancée ; il s'agissait, à partir d'un état de l'art, de concevoir des méthodes pour assurer au mieux le versionnage des concepts terminologiques. Enfin, en marge du projet ANR TerSan, deux nouveaux buts se sont ajoutés : concevoir et implémenter des méthodes de mises à jour intelligentes des ressources terminologiques mais également permettre, pour certaines d'entre elles, une interopérabilité particulière appelée « flux de terminologies ».
Tous ces objectifs tendent à se rejoindre vers un système unique agissant comme une plateforme permettant le stockage, l'édition, la mise à disposition et l'exploitation de terminologies faisant référence ; ce serveur, appelé S3M sera donc le socle de mes travaux mais aussi, nous le verrons plus tard, d'un système d'information entier.

1.5 Organisation du mémoire

Ce mémoire s'organise en sept chapitres. Deux chapitres sont dédiés au contexte de travail, aux objectifs et à l'état de l'art et définitions liés à cette étude. Trois chapitres développent les méthodologies et modèles conçus lors de ces travaux ainsi que leurs mises en œuvre. Enfin, un chapitre expose les résultats, les évaluations et les applications couplées à la plateforme 3M et le dernier chapitre conclut l'étude et énonce quelques perspectives.

Chapitre 2

État de l'art des classifications ou Systèmes d'Organisation de la Connaissance et définitions

Sommaire

Dans ce chapitre, il est question de définitions des concepts manipulés plus loin dans ce manuscrit. Ces définitions sont basées, d'une part sur un état de l'art du domaine mais également des domaines annexes, et d'autre part sur la vision que j'ai pu développer dans le cadre de ce projet de recherche. Nous allons tout d'abord présenter les différents types de classifications les plus utilisés puis montrer, via des exemples précis, comment et pourquoi ils sont utilisés.

2.1 Un peu d'Histoire

La création de classifications est essentiellement issue des sciences, et précisément des sciences du vivant. Conçues initialement pour consigner la connaissance, puis la partager, les classifications se sont révélées ensuite utiles pour établir des bases communes de désignation des concepts. Dans chaque discipline, un jargon devient peu à peu un champ lexical puis un lexique à part entière, qu'un ensemble d'utilisateurs partage pour parler la « même langue ». Cela facilite la pratique de la discipline, son enseignement et son utilisation.
L'exploitation plus poussée des classifications concerne la Recherche d'Information (RI) et d'autres traitements de l'information (statistiques, etc.). Dès le XIXème siècle, on a conscience que l'utilisation d'une classification pourrait améliorer la prise en charge des patients et la prévention des épidémies. En 1839, le docteur William Farr écrit : « *Les avantages d'une nomenclature statistique uniforme, même imparfaite, sont si évidents qu'il est surprenant qu'aucune attention n'ait été accordée à sa mise en vigueur dans les tables mortuaires. En de nombreuses circonstances, chaque maladie a été désignée par trois ou quatre termes, et chaque terme a été appliqué à de nombreuses maladies différentes : des noms vagues et impropres ont été employés, ou bien des complications ont été enregistrées à la place des maladies primitives. Dans ce domaine de la recherche, la nomenclature est d'une importance aussi grande que les poids et mesures dans les sciences physiques, et elle doit être établie sans délai.* » (traduit de l'anglais dans le *First annual report. London, Registrar General of England and Wales*). Un peu plus tard, on voit apparaitre des collections structurées de mots selon leur sens : elles sont nommées thésaurus par P.M. Roget dans le *Thesaurus of English Words and Phrases* [Roget, 1856]. En 1876, Melvil Dewey propose la première version de sa classification (*Classification Décimale de Dewey*) dédiée à classer les documents des bibliothèques.
Revenons à la médecine avec Jacques Bertillon qui, en 1893, intronise la *Classification des causes de décès*. En 1948, l'Organisation Mondiale de la Santé (OMS) reprend ses bases fondamentales pour créer la *Classification statistique Internationale des Maladies, traumatismes et causes de décès*, rebaptisée plus tard CIM (Classification Internationale des Maladies), qui est aujourd'hui toujours en vigueur avec ses révisions 9 et 10. Les terminologies vont ensuite se succéder dans tous les domaines

et leur utilisation va se systématiser.
C'est dans les années 1990 que le concept d'ontologie voit le jour. L'idée est de représenter la connaissance selon des langages formels via des outils mathématiques, sémantiques et informatiques.
Aujourd'hui, il existe toutes sortes de classifications, plus ou moins complexes ou volumineuses, dans des domaines extrêmement vastes ou très précis. Les Sciences en comptent sans doute le plus et leur gestion et leur utilisation deviennent de plus en plus indispensables dans les systèmes d'information.

2.2 Types de classification ou systèmes d'organisation des connaissances (SOC)

Les Systèmes d'Organisation des Connaissances (SOC) sont définis formellement comme l'ensemble des vocabulaires contrôlés [Hodge, 2000],[Binding & Tudhope, 2004]. Ils sont de différentes natures et ont différentes structures. Leurs buts sont très variés et il n'est pas exagéré d'affirmer qu'il existe autant d'objectifs de SOC que de SOC eux-mêmes. Cela se justifie simplement par le fait que chaque éditeur qui créé un SOC possède un objectif bien précis et très souvent orienté vers un cas d'usage spécifique. Tous les SOC n'ont pas forcément pour but initial d'organiser la connaissance mais de coder, d'annoter, de structurer l'information. La notion de « connaissance » est désormais très souvent une conséquence indirecte de la création d'un SOC.

L'utilisation des SOC

Dans la majorité des cas, la création d'un SOC part d'un besoin pour décrire des ressources. Les buts principaux d'utilisation des SOC sont les suivants :

- représenter les connaissances (stocker, décrire, structurer et annoter) d'un domaine spécifique ;
- rechercher l'information et effectuer des statistiques (indexation, codage) ;
- inférer des faits, des règles (raisonnements, subsomption, etc.).

Les SOC servent souvent à décrire des documents. Pour cela, on utilise 3 notions, suivant les domaines et les utilisations :

- l'indexation : il s'agit de repérer des termes représentatifs d'un document pour ensuite les réutiliser lors de la RI, tout comme un index à la fin d'un livre ;
- l'annotation : il s'agit d'ajouter un commentaire, contrôlé ou pas, à une portion de document. Dans certains cas, cela permet la RI mais ce n'est pas nécessairement le but recherché ;
- le codage : essentiellement utilisé dans le jargon des professionnels de santé, le codage consiste en l'ajout d'un code de SOC à un document (comptes-rendus,

images, etc.). Cela se rapproche donc de l'indexation si ce n'est que le but est avant tout socio-économique et non pas de RI.

Par abus de langage, ces 3 notions sont souvent confondues. Techniquement, cela importe peu car il s'agit de rattacher un concept de SOC à un document. Cependant, le but varie et influence fortement la façon dont sont choisis ces concepts.
Aujourd'hui, il existe un très grand nombre de SOC, particulièrement en Sciences. Beaucoup correspondent à une discipline ou à une sous-discipline particulière (médecine, cancérologie, biologie, anatomie, etc.) puisque l'un des principes de la création des SOC est l'établissement d'un vocabulaire contrôlé et standardisé.

2.3 Les classifications et leurs formes

Il existe de nombreuses classifications de termes également appelées « vocabulaires contrôlés ». Elles varient en taille, en complexité et en spécificité. Pour être plus juste, il faut dissocier le fond de la forme, à savoir le lexique de sa structure. Ainsi, un vocabulaire contrôlé correspond au lexique (aux termes à proprement parlé) alors qu'une classification est un système pour organiser ce lexique.
La forme la plus simple d'un SOC est une liste de termes. Un dictionnaire est un peu plus complexe car les termes sont triés et à chaque mot, il renvoie à une définition, à d'autres mots. Les SOC les plus utilisés sont les terminologies et les ontologies, dans l'ordre croissant de complexité de structure ou de formalisme (cf. figure 2.1). La plupart du temps, les SOC définissent des concepts, spécifiques d'un domaine, qui sont en relations entre eux [Duclos *et al.*, 2013]. Vanopstal dresse un état des lieux très intéressant sur la place des termes utilisés pour définir ces différentes classifications [Vanopstal *et al.*, 2011].

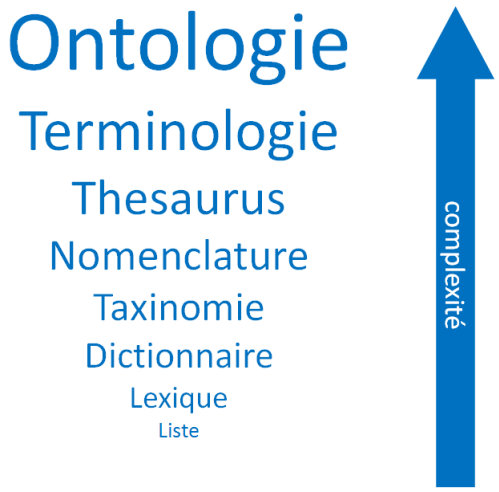

FIGURE 2.1 – Diversité et complexité des principaux types de SOC

2.3.1 Terminologie

Une terminologie est « une liste de termes d'un domaine ou d'un sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques, cette liste étant ou non structurée » [Lefèvre, 2000]. En d'autres mots, une terminologie implique la normalisation des termes d'un domaine afin de pouvoir les organiser les uns par rapport aux autres [Zweigenbaum, 1999]. Ces termes sont des entités et informations linguistiques qui peuvent s'associer entre elles. La structuration de ce lexique est très hétérogène d'une terminologie à une autre.

2.3.2 Thésaurus

Plus précis, il s'agit de terminologies dont les termes sont ordonnés et reliés entre eux par des relations hiérarchiques, d'équivalence ou d'association. Ce terme semble réservé aux terminologies documentaires : les termes, ou descripteurs, servent à indexer, ou décrire, des ressources documentaires.

2.3.3 Ontologie

La première définition de l'ontologie dans le domaine informatique est donnée par Gruber comme « *a specification of a conceptualization* » [Gruber, 1993]. Mais Guarino dénombre plus tard pas moins de 7 définitions du terme ontologie en informatique [Guarino & Giaretta, 1995]. Néanmoins, Schulz estime que la tendance actuelle est de considérer une ontologie comme une représentation des entités réelles [Schulz & Jansen, 2013]. Cette vision n'est pas partagée par tous et des définitions centrées sur l'interopérabilité ou l'utilisabilité sont aussi acceptées [Rzhetsky & Evans, 2011]. Il semble donc que cela dépende, comme l'affirmait Charlet [Charlet *et al.*, 2006], de la finalité de l'ontologie.
En pratique, il s'agit d'un ensemble de concepts et de relations dont la structure repose sur un formalisme standard. Le principe étant notamment d'être exploitables par des machines. Les algorithmes peuvent alors permettre d'effectuer des raisonnements sur ces concepts et même faire de l'inférence. Les termes contenus dans une ontologie ne sont plus les éléments centraux mais ils appartiennent à des entités plus larges que sont les concepts. Les ontologies permettent d'établir des règles et des contraintes sur les concepts et leurs relations.

2.3.4 Comparaison entre terminologie et ontologie

Lors de ces travaux de thèse, l'un des objectifs principaux était l'intégration de n'importe quelle nature de SOC, y compris les ontologies. Étant donné les choix méthodologiques et techniques conduisant à un modèle terminologique, il a fallu appréhender, comprendre et analyser les ontologies, autant sur le fond que sur la

forme. En plus d'une bibliographie et de l'étude sur la représentation formelle de la FMA (cf. 4.8.3), nous avons mené une étude sur la comparaison entre terminologie et ontologie. Il s'agit d'une problématique récurrente en ingénierie des connaissances et plus particulièrement en informatique médicale, tant l'utilisation de ces SOC est importante.

Il n'est pas possible de parler de la comparaison entre terminologie et ontologie sans mentionner un abus de langage extrêmement courant : pour beaucoup de scientifiques, les ontologies sont une forme structurée des terminologies, et ce, particulièrement chez les anglo-saxons. Un vocabulaire contrôlé contenu dans un format ontologique (OWL par exemple, cf. 2.8.4) ne peut pas être considéré comme une ontologie (confusion contenu/contenant). De mon expérience et celle de collaborateurs, à chaque fois qu'il est nécessaire de créer un vocabulaire contrôlé (pour un projet de recherche par exemple), le premier mot qui vient à l'esprit des gens est « ontologie ». Certains y mettent un vrai sens mais d'autres se trompent. La distinction entre les deux est bien réelle et créer une terminologie ou une ontologie sont deux choses assez éloignées tant leurs buts peuvent être différents.

Pour autant, terminologie et ontologie sont toutes deux définies comme des ensembles de termes ou concepts reliés entre eux et spécifiques d'un domaine. On peut parler de concept terminologique qui est la combinaison indissociable d'une dénomination et d'un concept. En outre, les ontologies sont structurées alors que les terminologies ne le sont pas toujours. Ces deux types de SOC définissent des propriétés aux concepts que sont les attributs et les relations. Les ontologies sont intrinsèquement constituées d'un modèle, plus ou moins formel. Les terminologies n'en ont pas toujours mais c'est souvent le cas, même si c'est implicite. Les ontologies sont plus complexes, non seulement à cause d'un modèle formel mais surtout à cause des différents axiomes applicables aux concepts et à leurs propriétés. On définit des règles précises sur les cardinalités, la symétrie, etc.

Concrètement, on peut dire que la terminologie s'intéresse plus au lexique contenu et aux notions (*terme* et *-logie* : « étude des termes » !) alors que l'ontologie se concentre plus sur la modélisation des concepts et sur leurs interactions. Roche [Roche, 2005] parle même d'une différence entre la pratique (pour la terminologie) et la métaphysique (pour l'ontologie).

Définir les différences de fond et de forme des terminologies et ontologies ne les oppose pas nécessairement. Elles ont des buts certes éloignés mais peuvent être employées à des fins complémentaires dans les systèmes d'informations (pratique) mais également en Sciences (théorie). Ainsi, comme conclut Grabar, dans son étude entre terminologie et ontologie [Grabar *et al.*, 2012], la plupart des spécialistes établissent un continuum entre terminologies et ontologies, et pas nécessairement une dichotomie.

Comme nous le verrons plus tard (cf. 4.8.3), l'intégration des ontologies (ou plutôt

des lexiques des ontologies) dans notre système repose sur une méthodologie simple et pragmatique.

2.4 Notion de concept

La notion de « concept » est équivalente que l'on parle d'ontologie ou de terminologie. Ce concept désigne une idée, une chose que ce soit donc abstrait ou physique. Il s'agit de quelque chose que l'on peut « concevoir » au sens intellectuel du terme. Idéalement, un concept est unique et doit avoir un et un seul sens, en tous cas, pour le concept terminologique ou ontologique. Il s'agit de l'essence même des vocabulaires contrôlés : il est nécessaire de désigner des concepts par leur nature unique et leur propriété à désigner quelque chose de précis et de commun à un domaine donné. Un concept ne doit pas être discuté au sein d'un SOC. Cela doit être clairement défini et sans équivoque.
Ainsi, quelque soit la culture, la langue ou même l'époque, un concept terminologique ou ontologique devra conserver son sens sous peine d'introduire de graves contre-sens au sein des SOC.
Ceci est extrêmement important car on peut déduire deux choses de cette invariance et de cette unicité des concepts :
- un concept est toujours le même quelque soit sa langue (ou sa traduction donc). En tous cas, le traducteur d'un libellé de concept devra absolument conserver le sens d'origine au moment de choisir la meilleure traduction possible.
- des relations entre concepts sont donc toujours vraies, quelque soient les libellés de ces concepts (ces relations peuvent tout de même s'avérer « fausses » si jamais la connaissance elle-même a changé : les « vérités » scientifiques sont parfois réfutées). En effet, si l'on change un libellé de concept et si l'on respecte l'intégrité du sens du concept, les relations avec d'autres concepts devraient rester vraies. Cependant, nombre de problèmes se posent quant au sens précis de chaque terme puisque chacun est à même de concevoir les choses différemment d'un autre [Gaudin, 1996].

2.4.1 La sémantique du concept

Si un concept de SOC désigne une idée bien précise, son libellé dit « préféré » (ou PT pour *Preferred Term*) peut évoluer avec le temps. En linguistique pure, si ce libellé évolue, il devrait évoluer vers un de ses synonymes, c'est-à-dire un terme qui possède le même sens. En toute théorie, si ces synonymes sont si identiques en sens avec un PT, ils ne devraient pas exister, c'est ce que l'on appelle des « synonymes parfaits » ou « parasynonymes ». D'ailleurs, nombre de linguistes et d'écrivains affirment qu'il n'existe pas de synonymes, comme Palmer en 1981 (cf. encart ci-contre) ou, bien avant cela, Flaubert dans la préface de *Pierre et Jean* de Maupassant : « Quel que soit [...] la chose qu'on veut dire, il n'y a qu'un mot pour l'exprimer, qu'un verbe pour l'animer, et qu'un adjectif pour la qualifier ». Encore avant, l'Abbé Girard consignait et décortiquait les différences entre mots considérés comme synonymes [Girard, 1769].

> « *It can, however, be maintained that there are no real synonyms, that no two words have exactly the same meaning. Indeed, it would seem unlikely that two words with exactly the same meaning would both survive in a language.* »
>
> [Palmer, 1981]

Dans l'univers des SOC, ou plus largement dans celui de la RI, ces nuances linguistiques sont peu exprimées voire inexistantes. Ceci pose donc des problèmes de sémantique pure. Il existe presque toujours des nuances entre un terme et ses synonymes (sauf pour les variantes orthographiques). Souvent, elles ne vont pas impacter le sens du concept mais parfois, selon le domaine, le contexte ou l'interprétation des auteurs, ce sens peut varier et ainsi changer le sens strict du concept. Modifier le PT d'un concept n'est donc pas anodin et engendre souvent des incohérences et peut aussi invalider les traductions et les relations sémantiques avec ce concept (par exemple, si le terme « Syndrome de Marfan » devient « Syndrome de Marfan type 1 », il devient plus précis et les relations d'équivalence avec d'autres concepts sont alors erronées).

Pour aller plus loin, dans les SOC, les synonymes sont souvent pris au sens large du terme et on ne distingue plus alors les synonymes « stricts » des « hyponymes » (sens plus précis) ou des « hyperonymes » (sens plus large). Pis, beaucoup de concepts possèdent des abréviations voire des sens dérivés plus souvent caractéristiques d'un « voir aussi » ou d'un « se référer à ».
La raison pour laquelle tous ces termes sont souvent regroupés en synonymes est un problème de RI : on peut admettre (voire très souvent s'attendre) d'avoir des résultats à la requête « ECG » traitant à la fois d'« Électrocardiographie », d'« Élec-

trocardiogramme » ou de « Cardioscope », même si, strictement parlant, ces termes ne sont pas synonymes et ne devraient donc pas être regroupés par le même concept. Un exemple classique en Santé est celui de la modélisation du MeSH avec ces Termes hyponymes, hypéronymes et reliés.

2.4.2 Les types de concepts

On définit un type de concepts de SOC comme étant un groupe ou une catégorie de concepts. Bien souvent, l'utilisation des SOC ou leur structure dépend de ces types. Ils permettent donc de créer des niveaux d'abstraction supplémentaires. Il convient cependant de préciser que la création de ces types n'est qu'une commodité intellectuelle et souvent technique pour organiser le SOC. Une autre manière de faire serait de n'avoir qu'un seul type de concepts et de typer chacun de ces concepts par un attribut particulier. Quoi qu'il en soit, les types de concepts sont très utilisés dans les SOC, en tous cas, dans les thésaurus et terminologies (exemples pour le MeSH des types de concepts `Descripteurs`, `Concepts Supplémentaires`, `Concepts`, `Termes`, `Types de publication` et `Qualificatifs`) ; on parle souvent de « schémas » pour désigner ces groupes de types de concepts.

2.4.3 De l'art des identifiants

Lorsque l'on crée un SOC et donc des concepts, il faut leur adjoindre un identifiant unique. Il existe, à notre connaissance, quatre façons de créer ces identifiants, classées par ordre croissant d'expressivité :

- par simple incrémentation : tous les concepts sont différentiés par un entier qui est incrémenté à chaque création. Dans ce cas, l'identifiant est inexpressif car il ne porte pas de sens particulier. Il s'agit d'un grand nombre de SOC (SNOMED CT, MedDRA, NCIT, FMA, ...) ;
- par incrémentation et type de concept : il s'agit encore d'incrémenter des entiers mais en y ajoutant un caractère (ou plusieurs) désignant le type de concept. Dans ce cas, l'identifiant est très peu expressif puisqu'il n'est possible que de déterminer à quel type il appartient sans connaître son PT. C'est le cas, par exemple, du MeSH, de HRDO ou encore de IUPAC ;
- par concaténation hiérarchique : pour un concept A créé, fils d'un concept B, son identifiant sera celui de B concaténé à un nouvel élément (chiffre ou lettre). Dans ce cas, l'identifiant est expressif puisqu'en connaissant le contexte (axe hiérarchique), il est possible d'en déduire un sens général. Il s'agit d'une bonne partie des SOC (SNOMED 3.5 int., CISP-2, ATC, CIM-10, ...) ;
- par combinaison de jeux de valeurs : les identifiants sont créés via un modèle précis. Ils sont souvent d'une longueur fixe où chaque caractère possède une signification particulière définie par des règles. Dans ce cas, les identifiants sont

dits très expressifs puisqu'il est possible d'en déduire un sens précis sans avoir à connaitre son PT. Très peu de SOC sont créés ainsi (CCAM et la terminologie d'interface de prescription du CHU de Rouen sont les seuls exemples dans le S3M).

Bien entendu, l'idéal serait de créer des identifiants les plus expressifs possibles. Cela permettrait aux utilisateurs de vite appréhender le SOC et de différencier rapidement les concepts, sans avoir à manipuler les PT. Cependant, cette méthodologie est coûteuse à mettre en place car il faut bien prévoir toutes les combinaisons possibles et garder une marge quant au nombre de caractères à manipuler ; les SOC évoluent rapidement et constamment. De plus, toutes les classifications ne sont pas compatibles avec une telle modélisation.

Par ailleurs, plusieurs SOC possèdent une particularité quant à la gestion de leurs identifiants. Par exemple, WHO-ART, SNOMED 3.5 internationale et MedDRA définissent certains concepts différents par un même identifiant ; dans ces cas, seul leur niveau dans la hiérarchie les différencie. Cela pose des problèmes évidents lors de l'intégration et de l'utilisation du SOC. Par exemple, les concepts WHO-ART « dermatite » (type de concept « Terme de Haut-Niveau ») et « dermatite allergique » (type de concept « Terme inclus ») ont tous deux pour identifiant `0007`. Les desiderata de Cimino mettent d'ailleurs en garde les éditeurs de SOCs sur ce sujet [Cimino & Zhu, 2006].

Enfin, à une dimension multi-terminologique, la redondance des identifiants inter-terminologiques est également possible. Plusieurs concepts de SOC différents peuvent très bien avoir le même identifiant. Afin de résoudre ce problème mais également celui de la redondance intra-terminologique, il faut mettre en place une méthode d'élaboration d'identifiant unique (cf. 3.5.1).

2.5 De l'art de la modélisation

Il n'est pas obligatoire de formaliser un modèle pour créer un SOC. Cependant, plus le niveau de complexité d'un SOC va être élevé plus il nécessitera un travail de modélisation poussé. Une simple de liste de termes non hiérarchisée peut très bien constituer un SOC valable et ne s'appuie sur aucun modèle réel. Cependant, lorsque l'on veut stocker un SOC dans un système de gestion de données, une modélisation devient *de facto* obligatoire. Lorsque l'on veut structurer un SOC et lui adjoindre ne serait-ce qu'une seule méta-donnée, la modélisation est également *de facto* créée ; un tableau contient forcément au moins une colonne ou au moins une ligne, un fichier XML contient au moins une balise ou une base de données contient au moins une table. Cela semble trivial mais la plupart des conceptions des SOC sur lesquels j'ai travaillé ont été des conceptions totalement informelles et peu réfléchies en tant que telles. Il est assez simple de remarquer que la structure des SOC dépend fortement

de sa date de création ; le format XML, par exemple, n'est apparu officiellement qu'en 1998[1] et la maîtrise d'un SGBD n'est pas connue de tous. À cette époque par exemple, peu d'éléments permettaient de créer des relations entre termes. Créer des SOC complexes via des outils informatiques peu adaptés présentaient un défi de taille. Par ailleurs, l'édition des SOC s'est souvent faite au fil de l'eau pour les SOC les plus anciens et ont dû faire face à l'apparition de nouveaux concepts, de nouvelles façons d'organiser la connaissance et de l'utiliser dans les systèmes. Un des cas les plus frappants est celui du MeSH : créé en 1962 [Rogers, 1963] par la NLM[2] (suite à la *Subject Heading Authority List* de 1954), ce SOC n'a cessé d'évoluer, non seulement dans son modèle mais également dans son utilisation. Sa structure originelle n'était constituée que de Descripteurs et de Qualificatifs. Par la suite, les Concepts Chimiques Supplémentaires sont apparus pour palier le manque de concepts spécialisés en chimie (renommés par la suite Concepts Supplémentaires (SC), incluant également des maladies rares). Ces SC ne sont pas hiérarchisés mais rattachés à des Descripteurs. Les notions de Concepts MeSH et Termes MeSH ont également largement complexifié le thésaurus à tel point que le moteur de recherche PubMed, qui s'appuie sur le MeSH, n'exploite pas toute la subtilité du modèle actuel. En effet, les experts de la NLM n'indexent pas avec les Termes ou les Concepts MeSH mais avec les Descripteurs et les SC uniquement. La recherche par Concept est possible mais ne pointe que sur le Descripteur correspondant, rendant le résultat de la requête plus bruyant que prévu [Darmoni *et al.*, 2012]. Une analyse plus poussée est détaillée en section 3.5.2.
De nos jours, la création d'un nouveau SOC est très souvent plus formellement décrite avec l'élaboration « consciente » d'un modèle. J'ai donc été confronté aux deux cas, avec une grande majorité de modèles non formellement décrits. Cela est non seulement une difficulté en soi soulève également des problèmes manifestes dans la façon dont certains SOC ont été créés (cas de la MedDRA notamment avec son modèle de concepts redondants [Merrill, 2008]).

2.6 Relations entre concepts

Au sens linguistique, il existe 6 grands types de relations sémantiques : 4 hiérarchisantes (hyperonymie/hyponymie, holonymie/méronymie) et 2 non-hiérarchisantes (synonymie, antonymie). Les sens sont identiques à ceux présentés dans la section 2.4.1 mais concernent, dans notre étude, des relations entre concepts et non entre termes.
Les relations hiérarchiques sont essentiellement divisées en deux catégories :

1. http://www.w3.org/TR/1998/REC-xml-19980210
2. *National Library of Medicine*, bibliothèque états-unienne spécialisée en médecine et pionnière dans la gestion des documents numériques et de leur classification

- les relations génériques ou *is-a*, qui permettent de définir des sous-types de concepts. Par exemple, la « grippe humaine » est une « maladie virale » ;
- les relations partitives ou *part-of*, qui permettent d'imbriquer les concepts les uns dans les autres (le tout et une partie du tout). Par exemple, une « phalange de doigt » est une partie d'« un doigt ».

Ce dernier type de relation peut par ailleurs prêter à confusion puisque *part-of* (« fait partie de ») peut être interprété de différentes manières, comme « est situé dans » (localisation spatiale) ou « est contenu dans » (sous-type) [Schulz *et al.*, 2006].
Par ailleurs, comme l'explique [Kister *et al.*, 2011], ces types de relations sont bien des constantes dans l'univers des SOC mais ne peuvent pas constituer l'ensemble des relations entre concepts. Pour cela, on définit des relations « conceptuelles » ou « riches » qui s'appliquent à deux concepts mais qui ne sont pas des relations de types sémantiques au sens propre. Par exemple, le concept FMA « muscle oblique supérieur » (49039) est en relation « est innervé par » avec le concept FMA « nerf trochléaire » (50865). D'autres exemples sont présentés en détails dans la section 4.6.3.
Les relations sémantiques et conceptuelles constituent un contenu riche au sein de chaque SOC et entre SOC et participent à la construction de réseaux, plus faciles à manipuler dans un univers multi-terminologique. Le réseau sémantique est donc défini comme l'ensemble des relations sémantiques reliant les concepts de SOC. Son intérêt n'est pas seulement d'apporter de la connaissance mais aussi de permettre l'interopérabilité des SOC et la création de flux d'informations (cf. section suivante).

Directions et domaines des relations

Les relations peuvent être de n'importe quelle nature et peuvent s'appliquer dans différentes directions selon cette nature. On définit la direction ou le sens d'une relation d'un concept vers un autre. Par exemple, une relation d'implication ne fonctionne souvent que dans un sens (si α implique β, β n'implique pas forcément α). Exemple : `la main fait partie du bras` mais `le bras ne fait pas partie de la main` : `le bras contient la main`. Ces deux relations (`fait partie` et `contient`) sont définies comme des relations inverses. De fait, une relation symétrique est son propre inverse (exemple : la relation `est égal à` est symétrique). La direction des relations est importante puisque pour chaque relation définie dans un modèle, il est important de réfléchir à ses deux directions potentielles.
De la même façon, on définit les domaines de relations qui sont les groupes de concepts sur lesquels peuvent s'appliquer certaines relations. Il s'agit alors de contraintes permettant de restreindre les sous-ensembles terminologiques impliqués de part et d'autre de chacune de ces relations. Exploiter ces domaines prévient d'éventuelles erreurs et assure donc la fiabilité des informations. Par exemple, dans le MeSH, la relation du type « Information d'indexation MeSH » ne peut relier d'un côté que

des Descripteurs MeSH et de l'autre des Concepts Supplémentaires MeSH.

2.7 L'interopérabilité, les alignements et le réseau sémantique

L'interopérabilité est définie comme la possibilité d'échanger l'information et l'utiliser entre différentes sources de données distribuées [Wegner, 1996]. Dans le cas d'un réseau sémantique de SOC interopérables, l'objectif de la plateforme 3M est de mutualiser les différentes sources d'interopérabilité au sens donc « sémantique », c'est-à-dire en conservant le sens des concepts associés.
Le terme « alignement » désigne la méthode permettant de mettre en correspondance deux concepts (terminologiques ou ontologiques) voire même deux SOC. On définit alors des relations d'alignement exacts qui sont des liens entre concepts considérés comme identiques mais appartenant à des SOC différents. L'interopérabilité sémantique s'établit donc grâce à ces alignements exacts. De nombreux travaux ont été menés pour créer cette interopérabilité permettant la transition d'un SOC à un autre. Parmi eux, on peut citer [Wang *et al.*, 2008], [Fung & Bodenreider, 2005] ou encore [Merabti *et al.*, 2012].
Le réseau sémantique, quant à lui, désigne l'ensemble des relations d'alignements permettant l'interopérabilité. En parcourant ce réseau, on peut notamment naviguer entre les SOC tout en gardant le sens. Passer d'un SOC à un autre est extrêmement utile car cela permet de bénéficier des spécificités de plusieurs SOC à la fois (contenu, structure, règles, etc.) et ainsi d'enrichir l'information. Permettre la création d'un réseau sémantique de qualité et le plus exhaustif possible est l'un des objectifs majeurs du serveur 3M.

Flux de terminologies

Pour aller plus loin dans l'idée d'un réseau sémantique multi-terminologique, la notion de flux terminologique a vu le jour dans le cadre d'interopérabilité entre SOC complémentaires. En effet, un contexte intéresse beaucoup les chercheurs en informatique médicale concernant la mise en correspondance de terminologies dites « locales », d'autres dites « d'interface » et des terminologies de référence.
Les SOC « locaux » sont, la plupart du temps, de simples nomenclatures ou des systèmes de codage utilisés par exemple dans établissement de soins. Ils sont dits « locaux » car leur portée s'arrête très souvent à un seul système d'information.
Les SOC de référence correspondent aux classifications reconnues et utilisées à des niveaux potentiellement élevés (national ou international).
Une terminologie « d'interface » est définie comme un SOC conçu spécialement pour un type d'utilisateur et un usage. Il s'agit en fait d'un lien entre l'humain et les terminologies locales et de référence. Le passage entre terminologie d'interface, terminolo-

gie locale et terminologie référentielle constitue typiquement un flux terminologique. Dans la réalité, il peut même s'intercaler d'autres vocabulaires plus spécifiques.
Les travaux de thèse de Nicolas Griffon [Griffon, 2013] traitent en profondeur de la particularité de ces différents niveaux terminologiques et des intérêts à les faire interopérer spécifiquement dans un flux. Dans ce cadre, nous pouvons citer [Griffon *et al.*, 2012] où il est proposée une méthode et un exemple pragmatique concernant la création d'une terminologie « d'interface » en imagerie médicale (voir Figure 2.2).
Les flux de terminologies posent un certain nombre de problèmes quant à la gestion des alignements, des versions des SOC mais également de leur utilisation. En effet, les terminologies de référence étant jugées peu applicables au codage (puisqu'elles ne sont pas utilisées en tant que terminologies locales), la mise en correspondance entre les différentes terminologies n'est pas simple : il existe des relations *1-n* (un code de terminologie d'interface peut correspondre à plusieurs codes de terminologie de référence par exemple), voire des relations *n-n*. Comme nous le verrons plus loin (cf. 3.6.2), l'implémentation de ce type de relations n'est pas triviale. En outre, la gestion des versions des alignements entre SOC est une problématique complexe, comme nous l'aborderons brièvement dans 4.5.2.
Enfin, il est important de noter que la création d'un flux terminologique ne s'effectue pas forcément avec le réseau sémantique proprement dit. Dans le cas de l'imagerie médicale, il s'agit plutôt de correspondances et non d'alignements. En effet, les concepts reliés ne sont pas considérés comme équivalents (en tous cas pas toujours). Il s'agit de relations conceptuelles typées comme décrites dans l'exemple suivant (Figure 2.3) : un médecin prescrit une glycémie pour un patient et voilà comment cela peut être transcodé via un flux de terminologies en biologie au CHU de Rouen.

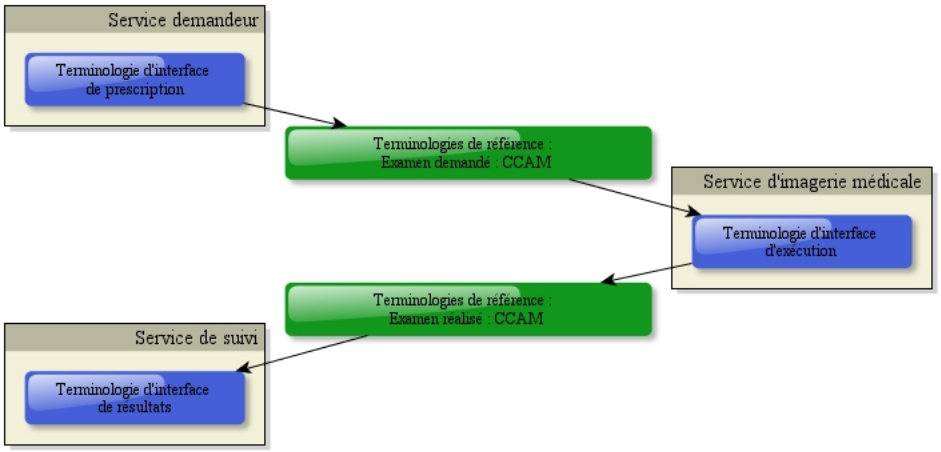

FIGURE 2.2 – Exemple de flux de terminologies en imagerie médicale [Griffon, 2013]

```
Prescription : Glycose Sang (BDXUR026)
=> Analyse (local) : Glycémie (BIOANAGLU)
=> Résultat (local) : Glucose (BIORESGLU)
=> LOINC (référence) : Glucose[Moles/Volume];Ser/Pla;Num (14749-6)
```

FIGURE 2.3 – Exemple de transcodage via un flux de terminologies en biologie pour une prescription de glycémie

Une des applications majeures des flux de terminologies consiste à permettre l'échange d'éléments de dossiers médicaux d'un établissement de soins à un autre sans perte d'informations et de contexte. Si les deux établissements possèdent des systèmes de codage internes différents, il peuvent tout de même effectuer un transcodage via une terminologie de référence et ainsi récupérer un code connu donc interprétable. C'est exactement la problématique du projet ANR TerSan [3] pour lequel le S3M a été mis à contribution.

La représentation des SOC, que ce soit pour l'interopérabilité, le lexique, voire même pour les modèles, peut s'avérer très hétérogène. Cela est essentiellement dû au fait que les SOC possèdent des typologies différentes mais également des formats de fichiers très variés. Il est important de bien les appréhender car notre approche nécessite une bonne connaissance des formats et des outils qui les exploitent. De plus, ils expliquent en partie la grande variabilité de structures des SOC.

2.8 Les formats

Les formats de fichiers ont évolué depuis les débuts de l'informatique. En effet, les besoins changent aussi bien techniques que fonctionnels : représenter la connaissance ne peut plus se faire par des simples tableaux à deux dimensions. La structuration des données permet une plus grande flexibilité des programmes les interprétant, une meilleure lisibilité pour les humains et une plus grande fiabilité (espaces mémoires restreints, sources d'erreurs diminuées, ...).
Les SOC intégrés dans le cadre de cette thèse ont des origines et des sources de données aux formats bien hétérogènes. Même si des standards ont été récemment définis, la plupart des SOC ont été créés avant leur établissement et transformer les fichiers électroniques originaux dans ces formats standards peut se révéler très complexe et très long.
Plusieurs formats classiques de fichiers de données sources sont détaillés ci-après, qu'ils soient standards ou non. Ils constituent aujourd'hui la vaste majorité des formats rencontrés dans l'univers des SOC en Santé et nécessitent parfois une prise en main délicate. Certains formats sont plus facilement lisibles par l'humain mais

3. Terminologies et Référentiels d'interopérabilité sémantique en Santé : ANR-11-TECS-0019, http://www.mondeca.com/fr/R-D/Projets/TerSan-Projet-ANR-TecSan-2012-2015

peu structuré et rigoureux, alors que d'autres sont difficilement directement interprétables par des humains car sont orientés machines.

2.8.1 XML

L'*eXtensible Markup Language* (XML) est défini comme un langage informatique de balisage. Les informations sont structurées par des balises qui peuvent être imbriquées les unes dans les autres. La liste des balises disponibles est extensible car il est possible d'en définir librement dans des espaces de noms. Les fichiers écrits en XML sont donc très pratiques puisqu'ils permettent la représentation de contenus complexes comme des arbres ou des tableaux à n dimensions.
Deux règles régissent les documents XML :

- ils sont définis via un schéma qui décrit formellement la structure des fichiers, avec la liste des balises dans leurs espaces de noms ;
- ils sont entièrement transformables vers d'autres documents XML (par exemple, via le langage XSLT [4]).

Le XML est un standard développé par le W3C [5] et connait un grand nombre de déclinaisons comme l'XHTML [6] ou le RSS [7]. Le XML est en soit un langage extrêmement bien adapté au stockage et à la représentation des connaissances. Des déclinaisons ont d'ailleurs été développées dans ce sens, comme le RDF/XML par exemple.

2.8.2 RDF/XML

Le RDF/XML est une syntaxe permettant de représenter des données structurées en RDF. Le *Resource Description Framework* (RDF), est un modèle de graphe conçu pour décrire formellement les ressources web et leurs méta-données. Ainsi, un document RDF est un ensemble de triplets qui sont définis de la sorte :

```
(sujet - prédicat - objet)
```

- le sujet représente la ressource (souvent par un identifiant unique appelé URI pour *Uniform Resource Identifier*) ;
- le prédicat est un type de propriété s'appliquant à cette ressource (souvent une méta-donnée) ;
- l'objet est la valeur de la propriété (donnée textuelle, nombre, date, ... ou bien une autre ressource).

4. *Extensible Stylesheet Language Transformations*, standard de transformation : http://www.w3.org/TR/xslt20/
5. *World Wide Web Consortium*, organisme de normalisation des technologies du web sur l'Internet : http://www.w3.org/
6. *Extensible HyperText Markup Language*, langage utilisé pour représenter les pages web
7. *Really Simple Syndication*, langage utilisé pour représenter la syndication de contenu web

Lorsque l'objet est une autre ressource, on considère le triplet comme une relation. Sinon, il s'agit d'un attribut.
Le RDF/XML est également un standard W3C depuis 2004 [8].

2.8.3 SKOS

Le *Simple Knowledge Organization System* (SKOS) [9] est un standard W3C depuis 2009 permettant de représenter des classifications, de la plus simple (liste) jusqu'à la terminologie et cela, notamment pour simplifier l'utilisation du OWL (cf. 2.8.4). Le SKOS s'appuie sur le RDF et définit en fait un ensemble de composants dédiés à la représentation des concepts terminologiques. Ainsi, les propriétés suivantes sont proposées pour les concepts :

- Termes Préférés (PT) ;
- Synonymes ;
- Définitions et notes ;
- Relations hiérarchiques ou associatives (dont celles d'alignements).

Il est également possible de regrouper les concepts dans des schémas de concepts.
La limite principale de ce langage est qu'il n'est pas possible de définir nativement des propriétés spécifiques à des concepts ou à des méta-données. Il faut, pour cela, étendre le SKOS à d'autres espaces de noms, ce qui n'est pas toujours considéré comme correct.

2.8.4 OWL

Le *Web Ontology Language* (OWL) est un langage de représentation des connaissances basé sur le RDF. Il s'agit d'un standard actuellement en version 2 W3C [10]. Le OWL est plus formel que le SKOS ou le RDF/XML puisqu'il permet de représenter précisément la connaissance via des fonctions (« constructeurs ») décrites dans le domaine de la logique de description. Il existe trois niveaux de formalisme à OWL correspondants à différentes logiques de descriptions plus ou moins complexes :

- OWL-Lite est la plus simple ;
- OWL-DL (pour OWL *Description Logics*) est une version décidable [11] de OWL et permet donc de garantir un résultat à des opérations d'inférence en un temps de calcul raisonnable (la logique de description est plus complexe) ;
- OWL-Full est la version la plus complexe et en fait indécidable ce qui en fait aujourd'hui une solution peu avantageuse.

Ainsi, OWL est le format le plus classique des ontologies, quelque soit le domaine d'application et est d'ailleurs apparu avant le SKOS.

8. http://www.w3.org/TR/2004/REC-rdf-syntax-grammar-20040210/
9. http://www.w3.org/2004/02/skos/
10. http://www.w3.org/TR/2012/REC-owl2-overview-20121211/
11. Que l'on peut résoudre de façon finie et déterminée

2.8.5 OBO

Par abus de langage, OBO (*Open Biomedical Ontologies*) désigne un format mais il s'agit à l'origine d'un projet de création de vocabulaires contrôlés partagés par les domaines de la médecine et de la biologie. Développé par un consortium états-unien, le format a en fait pour but de rendre les ontologies plus simples à parcourir, à lire pour l'humain, à étendre et il limite les répétitions [12].

2.8.6 Tableaux

Les formats de fichiers tabulaires sont très utilisés malgré leur manque d'expressivité et de standardisation. Les formats les plus classiques sont CSV [13], XLS [14] et ses variantes ou ODS [15].
Les tableaux sont des structures bidimensionnelles (la plupart du temps) et sont donc faciles à manipuler, autant par programme que par l'humain. C'est la raison pour laquelle ce type de format est encore largement répandu. En informatique médicale plus particulièrement, les tableaux constituent l'outil de prédilection d'un bon nombre d'utilisateurs.

2.8.7 Bases de données

Les Bases De Données (BDD) sont des conteneurs d'informations. Elles permettent de stocker des informations reliées entre elles et de les rechercher via des requêtes plus ou moins complexes. Ce type de système permet de structurer l'information et de la décrire au maximum.
Le terme « base de données » est souvent employé à la place de Système de Gestion de Base de Données (SGBD) qui correspond en fait au système qui manipule la BDD. Les SGBD sont classés dans plusieurs catégories, en fonction de leur typologie : relationnel, hiérarchique, orienté-objet, distribué, spatial, NoSQL, etc. La plupart des SGBD sont relationnels et constituent aujourd'hui les points centraux des Systèmes d'Information (SI) et ce, quelque soit le domaine. Depuis le début des années 2010 et l'essor des bases de données géantes, les technologies NoSQL (*Not Only SQL*) se sont développées et occupent désormais une place prépondérante dans l'univers des SGBD.
Quoi qu'il en soit, les formats de BDD sont nombreux puisqu'il existe beaucoup de SGBD sur le marché dont MySQL [16], Oracle [17], PostgreSQL [18], Microsoft Access [19].

12. http://www.geneontology.org/GO.format.obo-1_4.shtml
13. *Comma-Separated Values* : http://tools.ietf.org/html/rfc4180
14. Microsoft Excel
15. *OpenDocument Spreadsheet* de la suite Open Office
16. http://www.mysql.com/
17. http://www.oracle.com/
18. http://www.postgresql.org/
19. http://office.microsoft.com/fr-fr/access

Ces SGBD sont relationnels et sont fondés sur le SQL. Le *Structured Query Language* (SQL) est un langage informatique permettant d'exploiter les BDD en recherchant, ajoutant, modifiant ou supprimant des données du système.

2.9 Synthèse du chapitre

Dans ce chapitre, il a été question de définir les SOC, leurs différentes typologies mais aussi leurs propriétés. Nous avons vu qu'il en existait beaucoup, dans des formats hétérogènes, mais constituant tous des ressources de connaissances importantes. Concevoir et réaliser un système permettant de stocker plusieurs SOC simultanément, de les rendre interopérables, et ce, dans plusieurs langues constitue le cœur des travaux présentés ici.

Chapitre 3

L'approche multi-terminologique, multi-discipline, multi-lingue

Sommaire

Dans ce chapitre, je présenterai les différents modèles et applications existants proches de nos besoins à ce jour. Puis, je détaillerai les méthodes et outils de création d'un modèle (ou méta-modèle) permettant l'approche combinée multi-

terminologique, multi-discipline et multi-lingue. Ce chapitre se terminera par la présentation de ce modèle.

3.1 État de l'art

Plusieurs approches existent aujourd'hui pour stocker différents SOC dans un seul et même modèle afin de créer un serveur multi-termino ontologique. Même si souvent les buts diffèrent, les résultats sont très proches. J'ai étudié 4 grands systèmes, tous orientés Santé/Biologie : l'UMLS, BioPortal, le EBI Ontology Lookup Service et LexGrid/NCI Term Browser.

3.1.1 L'UMLS

En 1986, la NLM (National Library of Medecine) a lancé un programme de développement sur plusieurs années, nommé « Unified Medical Language System » (UMLS) [Humphreys *et al.*, 1998]. Ce projet associe plusieurs équipes de recherche et entreprises commerciales de différentes disciplines médicales ou d'informatique. Le but du projet UMLS est de fournir un accès intelligent aux ressources terminologiques biomédicales dispersées dans des bases de données multiples et disparates [Duclos *et al.*, 2013]. Une des caractéristiques de cette agrégation est de proposer un lien entre les différentes terminologies biomédicales intégrées. Par conséquent, l'un des objectifs de l'UMLS est de fournir une plate-forme permettant de regrouper tous les thésaurus, nomenclatures, ontologies et autres classifications existantes dans le domaine médical [Bodenreider, 2004] et le les relier entre eux.

Le méta-thésaurus

La caractéristique principale, qui est l'origine même de l'UMLS, est son méta-thésaurus. Le principe est de regrouper les classifications dans un seul et même système et d'associer à chaque concept un identifiant unique CUI (pour *Concept Unique Identifier*) : le sens d'un concept étant par essence conservé d'un SOC à un autre, des concepts de SOC différents peuvent donc partager le même CUI. En plus du CUI, d'autres identifiants uniques ont été créés pour gérer les lexiques comme le *Lexical Unique Identifier* (LUI) qui regroupent les variations lexicales pour un terme donné (uniquement en anglais). Les *String Unique Identifiers* (SUI) permettent quant à eux de distinguer les termes strictement identiques mais avec des sens différents (polysèmes). Enfin, les *Atom Unique Identifiers* (AUI) sont associés à chacun des termes de chacun des SOC (niveau le plus fin).
L'intérêt du méta-thésaurus, grâce aux CUI, est d'assurer non seulement une interopérabilité entre SOC mais également d'entretenir à un niveau supérieur de la

modélisation un système de connaissance unifié. Ainsi, le *Semantic Network* (Réseau Sémantique) de l'UMLS est défini au niveau des CUI. Il s'agit d'associer à chaque CUI un Type Sémantique qualifiant la nature du concept (maladie, enzyme, tissu, etc.). Les Types Sémantiques appartiennent à des Groupes Sémantiques, encore plus généraux (troubles, produits chimiques, anatomie, etc.). La richesse de ce réseau, ses applications et la grande volumétrie gérée par l'UMLS font aujourd'hui de ce système un outil indispensable dans la gestion des SOC dans les disciplines bio-médicales.

Modèle de données

L'UMLS repose sur une base de données relationnelle répartie en 11 entités pour les SOC et 13 entités pour les informations du méta-thésaurus. Il existe également un certain nombre d'index pour chaque langue.

3.1.2 BioPortal

Développé par le *National Center for Biomedical Ontology*[1] (NCBO) aux États-Unis, BioPortal est un entrepôt de SOC biomédicaux qui contient plus de 350 SOC dans différents formats [Noy *et al.*, 2009]. Il s'agit d'une librairie de SOC communautaires, conçue comme un « *one-stop shop* » (littéralement « guichet unique ») permettant aux utilisateurs de consulter, commenter et même d'ajouter eux-mêmes des SOC [Whetzel *et al.*, 2011].

Modèle de données

BioPortal a pour vocation de stocker des ontologies. Comme est discuté dans la partie 6.5.2, il ne s'agit en réalité que de formats ontologiques mais pas d'ontologies au sens formel. Les formats standards comme OWL ou OBO sont pris en compte nativement par le système de chargement de BioPortal pour les transformer en triplets RDF. L'entrepôt contient donc l'ensemble des SOC dans un format unique, sans méta-modèle particulier.

3.1.3 EBI OLS

L'outil Ontology Lookup Service (OLS) développé par l'Institut Européen de Bioinformatique (EBI - European Bioinformatics Institute) met à disposition plusieurs moyens pour requêter, consulter et naviguer parmi des ontologies et autres vocabulaires contrôlés biomédicaux [Côté *et al.*, 2006] [Côté *et al.*, 2010]. En plus de l'interface web dédiée (https://www.ebi.ac.uk/ontology-lookup/), un Service Web permet également d'accéder à ces fonctionnalités. Développée et conçue par

1. http://www.bioontology.org/

des bioinformaticiens, cette plateforme est avant tout utile pour les biologistes souhaitant annoter leur matériel (gènes, protéines, voies métaboliques, etc.). Ainsi, on y trouve essentiellement des SOC orientés par organisme ou par fonction biologique.

Modèle de données

Le modèle de OLS est inspiré de la partie « Ontologies » du schéma de base de données BioSQL [2]. Il s'agit donc d'une base de données relationnelle essentiellement tournée vers la définition des termes et relations entre ces termes. Cette base de données est alimentée automatiquement par des fichiers au format OBO.

3.1.4 LexGrid et le NCI Term Browser

LexGrid [3] (pour *Lexical Grid* ou « Grille lexicale » littéralement) est une suite d'outils permettant de stocker, rechercher et utiliser des SOC dans différents buts [Pathak *et al.*, 2009]. Développé par la Mayo Clinic [4], LexGrid propose des services d'accès aux SOC par l'intermédiaire des standards HL7 [5] CTS2 [6] (pour *Common Terminology Services* version 2) et d'une API dédiée (LexBIG). Plus précisément, CTS2 définit une syntaxe et des méthodes primitives pour manipuler des terminologies. Le principe est d'instancier des clients CTS2 dans plusieurs systèmes distants afin qu'ils puissent interagir sans ambiguïté.
L'implémentation principale de LexGrid est LexEVS qui est le serveur terminologique central contenant les principaux SOC en Santé utiles aux projets de la Mayo Clinic. En outre, il existe un portail appelé *NCI Term Browser*[7] permettant de consulter et de rechercher des concepts terminologiques d'une vingtaine de SOC dans LexEVS.

Modèle de données

LexGrid est basé sur un méta-modèle décrivant les données minimales et la façon dont elles doivent être représentées quelque soit le vocabulaire contrôlé intégré. Représenté en XML, ce modèle est compatible avec la plupart des fonctions ontologiques. Très complexe, il nécessite une prise en main difficile compte-tenu de ses nombreuses ramifications et spécificités relatives aux standards HL7.

2. Open Biological Database Access (OBDA) - http://obda.open-bio.org/
3. https://wiki.nci.nih.gov/display/LexEVS/LexGrid
4. Fondation à but non lucratif implantée essentiellement aux États-Unis et spécialisée dans de nombreux domaines autour de la médecine : http://www.mayoclinic.org/
5. *Health Level 7*, organisation à but non lucratif ayant pour objectifs de développer des standards pour l'interopérabilité des systèmes de soins dans le monde : http://www.hl7.org/
6. http://informatics.mayo.edu/cts2/index.php/Main_Page, http://www.omg.org/spec/CTS2/1.0/
7. http://nciterms.nci.nih.gov/

3.2 Multi-terminologie et méta-modèle

Notre approche de modélisation d'un serveur 3M repose sur la création d'un méta-modèle terminologique, capable d'agréger n'importe quel type de données terminologiques. Il s'agit d'un socle commun à tout SOC intégré puisque chacun de ces SOC devra posséder un modèle terminologique héritant du méta-modèle présenté ici.

À l'issue du projet InterSTIS, le modèle 1M se présente ainsi (figure 3.1). En résumé, un concept terminologique doit être unique et posséder obligatoirement un identifiant ainsi qu'un type de concept (entités `TB_DESCRIPTEUR` et `TB_DESCRIPTEUR_TYPE_DESCR`). Il appartient à une et une seule terminologie (entité `TB_DESCRIPTEUR_TERMINOLOGIE`). En outre, il possède au moins un libellé préféré (PT pour Preferred Term en anglais), soit en français, soit en anglais, voire les deux. Par ailleurs, chaque concept peut présenter un certain nombre d'attributs (synonymes, liens hypertextes, images, définitions, etc.) (entités `TB_DESCRIPTEUR_UF`, `TB_DESCRIPTEUR_DEFINITION` et `TB_DESCRIPTEUR_ATTRIBUT`). Les concepts peuvent être liés entre eux par deux grands types de relations (entités `TB_DESCRIPTEUR_RELATION` et `TB_DESCRIPTEUR_BT_NT`) :

- Les relations intra-terminologiques qui impliquent deux concepts issus du même SOC. Les relations hiérarchiques (is-a ou part-of) permettant de structurer et/ou organiser les concepts d'un SOC font partie de ces relations intra-terminologiques.
- Les relations inter-terminologiques qui impliquent deux concepts issus de SOC différents. La plupart du temps, ces relations sont définies pour permettre l'interopérabilité des SOC (cf. 2.7) mais peuvent également résulter d'une complémentarité entre SOC pour décrire la connaissance (cf. 4.6.3).

Ce modèle 1M bilingue a permis la création du Portail Terminologique de Santé (5.1).

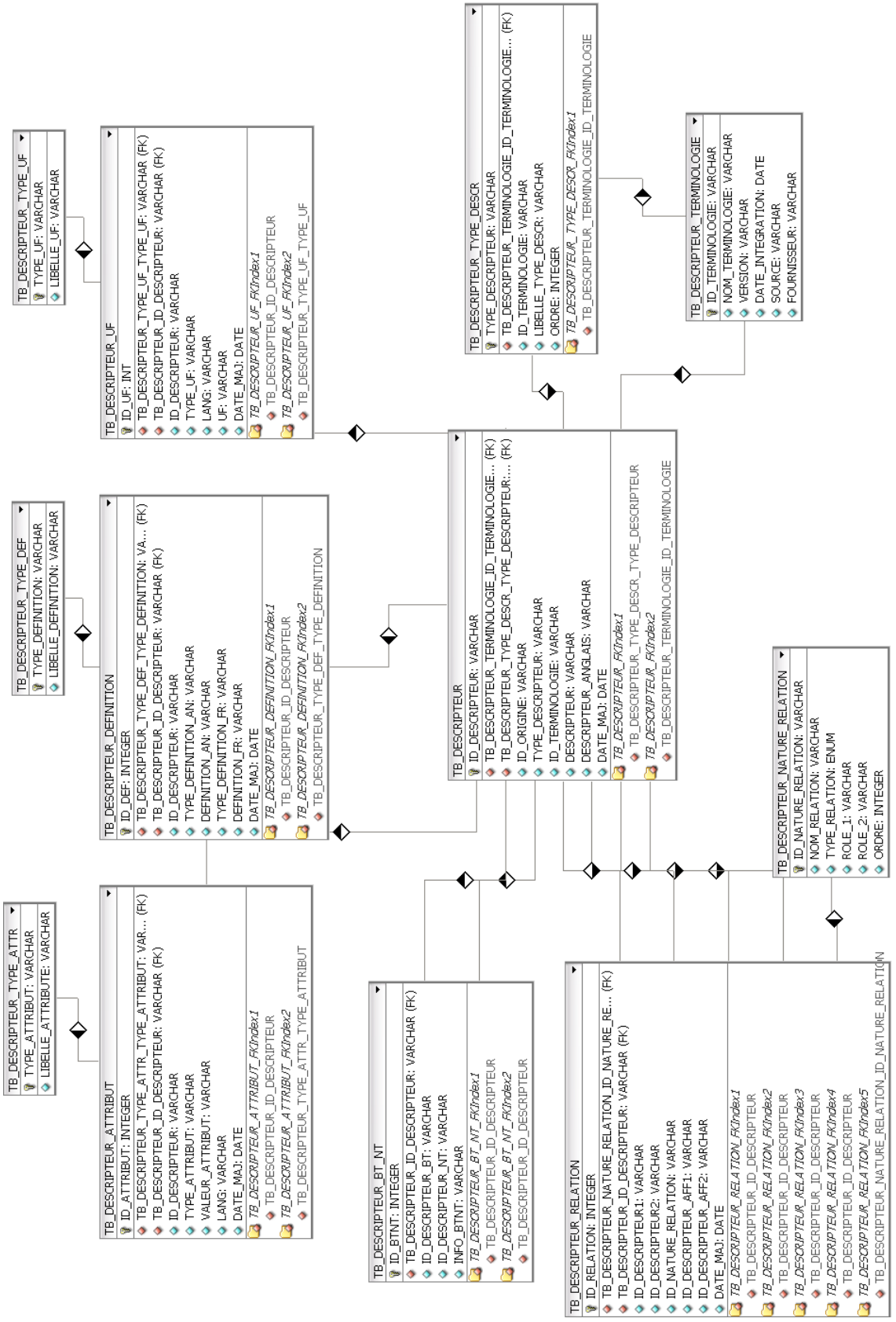

FIGURE 3.1 – Modèle Logique de Données Multi-Terminologique

3.3 Modèle multi-discipline

L'informatique médicale joue un rôle prépondérant dans ces travaux de thèse puisque l'équipe CISMeF travaille au quotidien à décrire ses ressources. Par ailleurs, les projets de recherche présentés précédemment sont fortement axés sur les domaines de la Santé et particulièrement sur la médecine. C'est la raison pour laquelle la majorité des exemples concernent des concepts médicaux (maladies, signes, dispositifs médicaux, etc.) ou des applications axées sur la Santé. Le but du deuxième « M » du modèle 3M est de démontrer que ce modèle est multi-discipline, et peut donc intégrer en son sein des SOC issus d'autres domaines. Nous verrons plus loin que son implémentation a été faite en matière de preuve de concept et que le nombre de SOC hors Santé ne constitue qu'une petite part de l'ensemble des SOC intégrés.

3.4 Modèle inter-lingue

L'objectif du troisième et dernier « M » était de permettre l'intégration de n'importe quelle langue pour n'importe quel SOC. En effet, les SOC étant très souvent spécifiques à un domaine, les termes complexes sont difficiles à appréhender dans une langue non maternelle. De plus, beaucoup de systèmes ont été conçus pour des usages particuliers afin de répondre à un besoin précis. Dans les décennies 1980 et 1990, la plupart des programmes ne sont pas encore développés selon des critères de multi-linguisme ou « d'internationalisation ». Ainsi, lorsque, par exemple, l'Organisation Mondiale de la Santé (OMS) édite la Classification des Maladies (CIM) en anglais, les institutions de chaque pays la reprend en la traduisant. Cependant, aucune règle ni aucune validation ne seront apportées pour uniformiser cela ou même constituer une CIM internationale unifiée.
Les SOC sont nativement construits autour des concepts et gérer les différentes langues d'un même terme n'est pas forcément complexe puisque l'on connaît, *a priori*, les correspondances entre termes de différentes langues car elles sont fournies directement par les éditeurs de SOC.
La représentation d'un terme en plusieurs langues se fait donc naturellement par le triplet « identifiant » - « libellé » - « langue ». Techniquement, il serait peu envisageable d'avoir une table avec autant de colonnes que de langues disponibles. C'est la raison pour laquelle il vaut mieux stocker autant de lignes qu'il y a de traductions. Nous développerons le modèle physique dans le chapitre 3.6. Conceptuellement, l'adaptation du modèle en inter-linguisme, et même multi-linguisme s'est fait aisément, toute la complexité du travail se reportant sur le modèle physique.

3.4.1 Inter-linguisme et multi-linguisme

Il ne faut pas confondre inter-linguisme (en anglais *cross-lingual*) qui désigne le fait de passer d'une langue à une autre en conservant le sens (au maximum) avec le multi-linguisme qui correspond au fait de gérer plusieurs langues dans un système donné [8]. Ainsi, le méta-modèle présenté dans ces travaux sera à la fois inter-lingue, car on pourra passer d'une langue à une autre via les concepts, mais il sera également multi-lingue, car on pourra rechercher des termes dans plusieurs langues à la fois.

3.4.2 Problèmes liés aux traductions

Étant donné que les SOC sont souvent développés dans une langue (très souvent l'anglais), la traduction se fait la plupart du temps *a posteriori*, par d'autres entités que l'éditeur original. Outre les erreurs de traduction (par exemple évaluées à 9.5% dans la version française de la FMA [Merabti *et al.*, 2011]), un autre type d'erreur peut s'avérer très problématique : certains libellés de concepts dépendent d'un contexte, c'est-à-dire de leur place dans la hiérarchie. En effet, lors d'une traduction « à la volée », il peut arriver que certains traducteurs éludent une partie du libellé qui se répète dans tous les concepts d'un même contexte. Ainsi, si par exemple, le concept « Cancer de l'œsophage » fait partie dans la hiérarchie du concept « Cancers », il peut arriver que la traduction deviennent « Œsophage » et non « Esophagus cancer » (si l'on traduit ici en anglais). Ce genre de raccourci peut s'avérer extrêmement dommageable car les traitements automatiques (RI, alignements basés sur du Traitement Automatique de la Langue, etc.) ne détecteront pas cette subtilité.
La traduction des SOC référentiels est devenue de plus en plus courante à l'ère de l'interopérabilité des systèmes. Ainsi, certains éditeurs définissent des guides de bonnes pratiques pour traduire leurs vocabulaires. On peut citer celui de la SNOMED CT [9], fourni par l'éditeur de ce SOC, la IHTSDO (*International Health Terminology Standards Development Organisation*). D'autres éditeurs proposent des outils dédiés pour faciliter le travail et minimiser les erreurs ; c'est le cas par exemple du *MeSH Translation Maintenance System* [Nelson *et al.*, 2004].
Nous verrons, dans la discussion, que ces problèmes de traduction peuvent avoir un impact très important sur un serveur multi-terminologique fortement dépendant de l'interopérabilité (cf. 6.2.2).

8. Le titre de cette thèse fait uniquement mention de Multi-linguisme. Ceci n'est que par pure commodité afin de conserver la notation 3M, plus pratique. Dans ces travaux, les deux notions sont souvent confondues, par abus de langage.

9. http://www.ihtsdo.org/fileadmin/user_upload/Docs_01/Publications/Quality_Assurance/IHTSDO_Translation_Guidelines_v2_02_20121211.pdf

3.4.3 Inter- et multi-linguisme et synonymie

Les SOC inter- et multi-lingues sont considérés comme tels car les concepts possèdent plusieurs libellés avec des langues différentes. La plupart du temps, les PT sont traduits d'une langue à une autre et l'équivalence entre les termes est évidente. Cependant, la gestion du multi-linguisme avec les synonymes est beaucoup plus floue. Dans tous les SOC que j'ai pu intégrer dans le 3M, aucun d'entre eux ne pouvait faire la correspondance exacte entre synonymes traduits. En d'autres termes, pour deux synonymes de langues différentes d'un même concept, il est impossible de savoir s'ils sont la traduction l'un de l'autre.
L'une des raisons de ce problème est qu'il n'existe pas de façon de modéliser cela dans les formats termino-ontologiques. Qu'il s'agisse du SKOS ou du OBO, par exemple, les attributs de synonymie sont multi-valuables donc savoir si tel synonyme est la traduction de tel synonyme est impossible sans élément supplémentaire. Une solution serait de créer un type de concept extrêmement précis (« Terme ») à chaque SOC et donc que chaque Terme ne possède qu'un seul libellé dans une langue (vision « atomique »). Cela est très difficile à mettre en place car très fastidieux, ne serait-ce que pour reprendre tous les synonymes de SOC déjà créés. Par ailleurs, les éditeurs n'apportent que peu d'intérêt à la représentation des synonymes. La plupart du temps, ces synonymes (qui n'en sont pas toujours, cf. 2.4.1) servent surtout à la RI. Cela constitue un manque notable du point de vue connaissance et linguistique ; si tous les synonymes avaient des traductions un à un, il serait possible de construire un dictionnaire d'une grande qualité.

3.5 Présentation du méta-modèle 3M

Le méta-modèle 3M encapsule tous les modèles de SOC que l'on souhaite y intégrer. L'entité principale, nommée `Descriptor`, correspond aux concepts des SOC. Chaque `Descriptor` appartient à une entité `Descriptor_type` (type de concept) qui elle-même appartient à une entité `Terminology`, qui correspond aux SOC. Enfin, les `Descriptors` possèdent un certain nombre de propriétés communes (attributs et relations). La Figure 3.2 illustre l'ensemble du méta-modèle 3M.
Voici la liste détaillée des attributs :

- RDF_RESOURCE : identifiant unique du concept dans le S3M (cf. 3.5.1) ;
- TYPE_ID : type de concept (`Descriptor_type`) du concept terminologique, ce type appartenant à une et une seule `Terminology` ;
- ORIGIN_ID : identifiant d'origine du concept dans son SOC natif ;
- RDFS_LABEL : libellé préféré (PT) du concept ;
- CISMEF_SYN : synonyme CISMeF ; tous les SOC sont susceptibles d'être enrichis par les experts de l'équipe afin d'améliorer son contenu (connaissance) et sa portée (RI) ;

- CISMEF_ACR : acronyme CISMeF ; pour les mêmes raisons que les synonymes ;
- CISMEF_DEF : définition CISMeF ; pour les mêmes raison que les synonymes ;
- COMM_SYN : synonyme de la communauté ; idem que pour les synonymes CISMeF mais cette fois la source du contenu est la communauté. Les utilisateurs peuvent ajouter des synonymes via l'application HeTOP (cf. 5.2.8) ;
- UMLS_CUI : les CUI de l'UMLS ; étant donné que beaucoup de SOC intégrés au S3M appartiennent à l'UMLS et que l'interopérabilité peut permettre le partage de ces identifiants de concepts, cet attribut a été défini au niveau du méta-modèle ;
- IS_OBSOLETE : booléen permettant de définir si un concept terminologique est désigné comme obsolète, c'est-à-dire plus utilisé. Cela est utile dans le cas du versionnage pour maintenir un suivi et une rétro-compatibilité (cf. 4.5).

Les `Descriptors` peuvent être liés entre-eux par des relations, intra- et/ou inter-SOC. On dissocie les relations hiérarchiques (`is-a` ou `part-of`) et les relations d'interopérabilité (alignements) des autres types de relations.
Il s'agit ici du méta-modèle 3M, c'est-à-dire des propriétés communes à tout SOC intégré dans le S3M. Chacun de ces SOC pourra bien sûr définir ses propres propriétés.

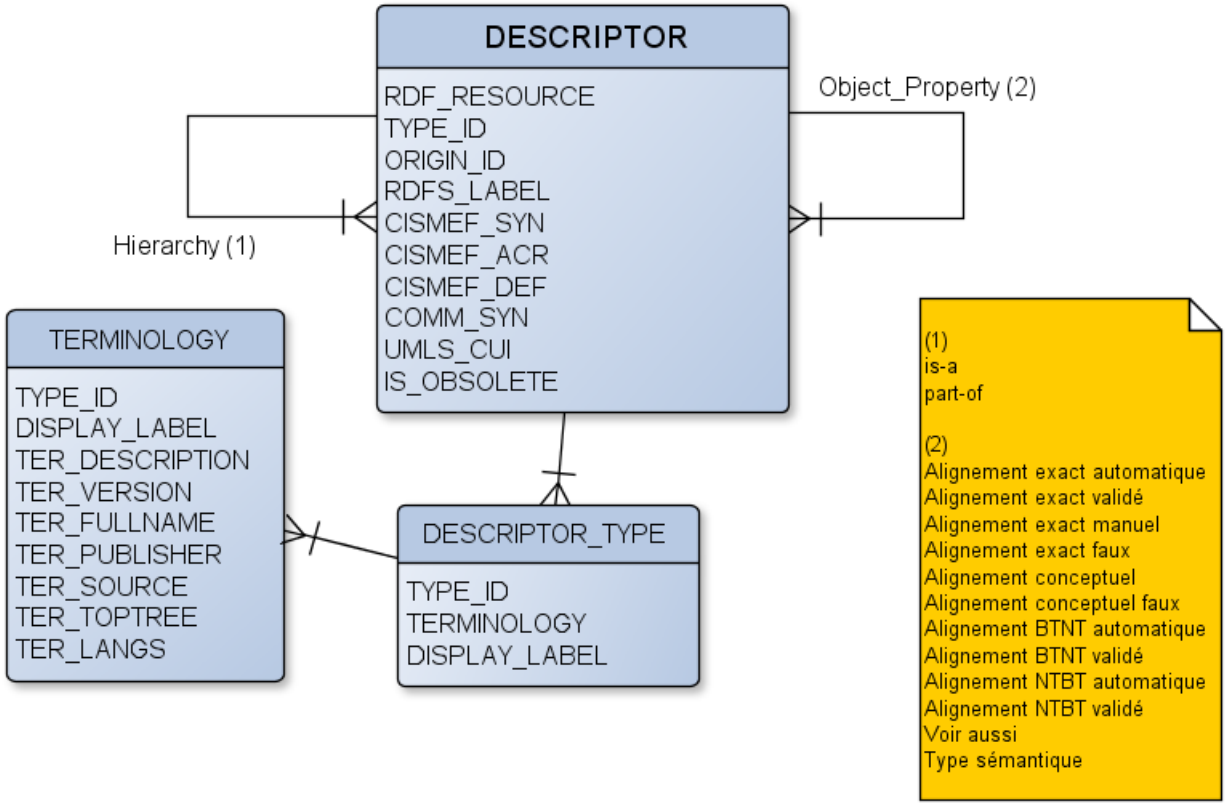

FIGURE 3.2 – Schéma conceptuel du méta-modèle 3M

3.5.1 Gestion des identifiants en multi-terminologie

Les identifiants de concepts doivent être uniques pour éviter toute incohérence. Or, il existe des cas de redondance d'identifiant entre SOC mais également au sein d'un même SOC (cf. 2.4.3). Pour palier ce problème, un système de normalisation d'identifiant de concept a été mis en place dans le S3M. La méthode consiste à concaténer 3 composants définissant ainsi de manière unique chaque concept :

- l'abréviation du SOC (exemple : CIM pour CIM-10) ;
- l'abréviation du type de concept du SOC (exemple : CA pour Catégorie CIM-10) ;
- l'identifiant d'origine du concept (exemple : J03 pour la Catégorie CIM-10 « amygdalite aiguë »).

Dans notre exemple, l'identifiant unique (« RDF_RESOURCE ») sera CIM_CA_J03.
Pour le cas particulier d'une redondance intra-terminologique, les identifiants des concepts « dermatite » et « dermatite allergique » dans WHO-ART seront respectivement ART_HLT_0007 et ART_IT_0007, par exemple.
Cette granularité suffit à désigner de façon unique chaque concept dans l'environnement du S3M.

3.5.2 Un pivot : le MeSH

Comme expliqué avant dans ce mémoire, le MeSH est une classification un peu à part dans l'univers des SOC en Santé. Il s'agit du plus ancien et sans doute du plus partagé des SOC à ce jour, il s'agit d'un pionnier également dans les SI documentaires [Lipscomb, 2000]. C'est pour cela que le MeSH est aujourd'hui un pivot dans une dimension multi-terminologique. De la même façon, puisque le MeSH est la classification historiquement utilisée dans CISMeF, il est utilisé *de facto* comme pivot. Pendant plus de 15 ans, les experts du CISMeF ont ajouté du contenu au MeSH, principalement en français (cf. 6.4.2). De plus, la nature même du MeSH, qui se veut généraliste en médecine et en général sur la Santé, en fait aujourd'hui une ressource incontournable en terme de richesse de contenu et d'exactitude terminologique. Le MeSH est actualisé en continu avec une grande mise à jour annuelle gérée, pour le français, par l'équipe DISC [10] (INSERM).
Pour toutes ces raisons, le MeSH est un pivot dans le S3M développé ici. Les premiers alignements constituant le réseau sémantique du S3M ont été effectués sur le MeSH. La majorité des travaux consistant à améliorer le contenu des SOC a impacté le MeSH en premier lieu (icônes VCM, intégration du réseau sémantique de l'UMLS, etc. : cf. 4.6).
Mais le MeSH n'est pas seulement un pivot en terme de contenu. Sa structure,

10. Information Scientifique et Communication : www.inserm.fr/qu-est-ce-que-l-inserm/organigramme/departements-et-services/information-scientifique-et-communication-disc

complexe et assez inédite dans l'univers des SOC, en fait une classification à part et pionnière dans la façon de modéliser un SOC, de l'utiliser et de représenter sa hiérarchie.

Un modèle particulier

Le modèle du thésaurus MeSH est en effet très particulier car il est conçu pour indexer finement des documents. Ainsi, le premier type de concepts du MeSH est nommé « Descriptor » (ou « Descripteur » en français) : il s'agit de plus de 25 000 concepts répartis en 16 axes couvrant un maximum le champ de l'indexation documentaire en Santé. Comme mentionné dans 2.5, le modèle du MeSH a évolué au cours du temps et offre quelques subtilités avec les d'abord les « Qualifiers » (ou « Qualificatifs » en français, puis les « Concepts Supplémentaires » et enfin les « Termes » et « Concepts » MeSH. Outre les problèmes qu'ils posent, ils constituent en fait une modélisation intéressante à analyser.
Tout d'abord, le MeSH permet d'effectuer la post-coordination (on parle « d'affiliation » dans un thésaurus), c'est-à-dire qu'il est possible de créer des combinaisons autorisées de Descripteur/Qualificatifs pouvant en fait être considérées comme des pseudo-concepts. Par exemple, si un document traite de la thérapie de l'asthme, il faut utiliser le « Descripteur » asthme affilié du « Qualificatif » thérapie. Du côté de la RI, la recherche « asthme/thérapie » pourra redonner tous les documents indexés de cette manière. Cette fonction est très pratique et extrêmement utilisée dans le MeSH, contrairement aux autres SOC. La post-coordination existe également dans quelques autres SOC depuis (SNOMED int. et CT, par exemple).
Ensuite, le MeSH possède un modèle imbriqué avec les Termes (plus bas niveaux) compris dans des Concepts, qui eux-mêmes sont rattachés à des « Records » (Descripteur ou Concept Supplémentaire). Cette structure, avant tout historique, est restée pour des raisons de facilité d'indexation et de RI essentiellement. Le vocabulaire du MeSH est désormais trop riche et volumineux pour être appréhendé en entier par les experts indexeurs (plusieurs centaines de milliers de termes). Ainsi, ces différents niveaux ont été créés pour minimiser cette complexité et raisonner uniquement au niveau des Descripteurs et des Concepts Supplémentaires. L'indexation est certes moins fine et la RI fonctionne toujours. Ainsi, comme montré dans les travaux de [Darmoni *et al.*, 2012], même si les éditeurs du MeSH n'indexent pas avec les Concepts (et *a fortiori* pas avec les Termes), il est tout à fait possible de le faire afin d'améliorer la précision.
Enfin, le MeSH définit des « Règles d'Indexation » précisant qu'en plus d'indexer avec tel Concept Supplémentaire, il faut également indexer avec tel(s) Descripteur(s). Cela engendre des règles métiers spécifiques à implémenter dans les systèmes pour parcourir les relations et inférer de nouvelles indexations.

Dans le S3M, le modèle du MeSH est un des rares modèles de SOC a avoir été modifié par rapport à l'original. En effet, les Termes, jugés peu expressifs et redondants dans la définition des Concepts, ont été agrégés avec ces derniers ; chaque Record possède au moins un Concept dit « Préféré » et éventuellement d'autres Concepts qui peuvent être plus ou moins représentatifs du Record. Ceux-ci constituent en fait un « sac » de termes synonymes du Record. Cela dit, il ne s'agit pas toujours de synonymes mais souvent d'hyponymes, d'hypéronymes ou de termes reliés. L'exemple du Descripteur MeSH « agueusie » (D000370) montré en Figure 3.3 illustre cela avec quatre Concepts reliés. Un PT « agueusie », un NT « agueusie hystérique », un BT « hypogueusie » et un RT « cécité gustative ». On voit bien ici que cette représentation entre Record et Concept MeSH n'est pas commode puisqu'elle duplique des termes et apporte une grande imprécision quant à la gestion de la synonymie.
Le dernier changement de modèle concerne la branche « caractéristiques d'une publication » qui a été déplacée vers un nouveau type de concept « Type de publication » (à la place de Descripteur) pour permettre une qualification plus précise des documents (types de ressources) et corriger la modélisation étonnante du MeSH qui consiste à créer un Descripteur pour un type de ressource (exemple : « Étude d'observation » (D064888)) et un autre pour son utilisation (« Étude d'observation comme sujet (D064887) »).
Le modèle conceptuel simplifié du MeSH dans le S3M est montré en Figure 3.4.

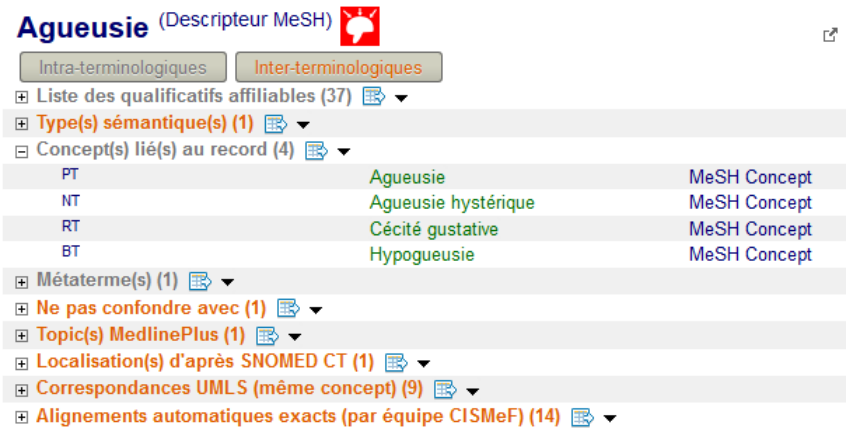

FIGURE 3.3 – Exemple de relations entre Descripteur MeSH et Concepts MeSH dans HeTOP

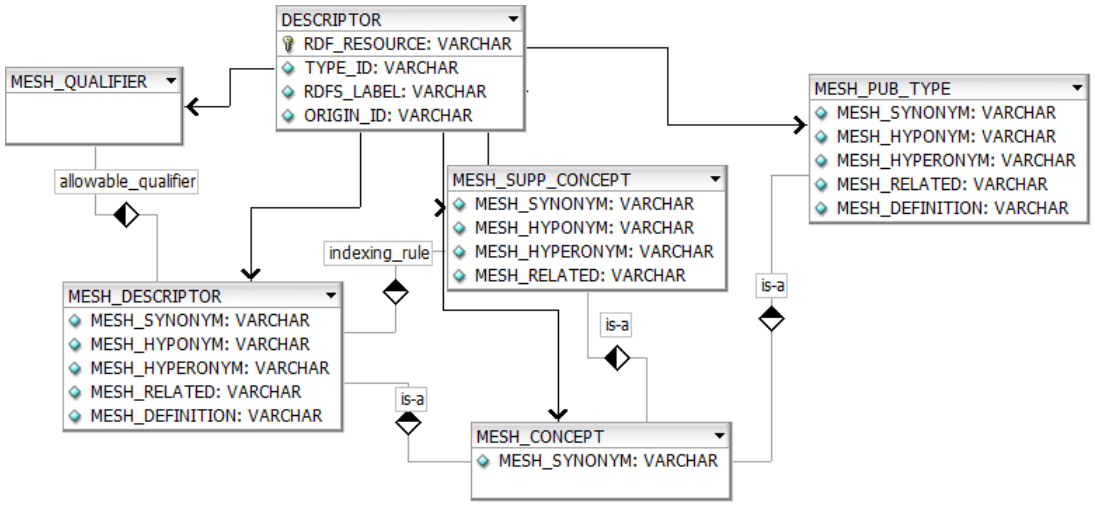

FIGURE 3.4 – Schéma conceptuel simplifié du MeSH dans le S3M

3.6 Modèle logique de données générique

Les modèles de base de données relationnelles sont souvent basés sur la méthode Merise [Tardieu *et al.*, 1995] ou le langage UML (Unified Modeling Language) [Soutou, 2002]. Ces outils sont similaires à la notion de Programmation Orientée Objet (POO). En effet, chaque « entité » correspondra à une table dans la base de données et sera, la plupart du temps, le reflet d'un objet physique ou théorique du monde réel (un patient, un document, un pays, etc.). En conséquence, il y aura autant de tables que d'objets à gérer dans la base. De plus, d'autres tables (systèmes, jointures, etc.) pourront s'ajouter à la liste si bien que le nombre de tables pourra atteindre plusieurs dizaines voire centaines dans des systèmes d'information complexes. Étant donné que les colonnes des tables correspondent, le plus souvent, aux attributs d'une entité, il faut donc également connaître tous ces attributs et lors d'évolutions d'un système, il est toujours complexe d'ajouter ou retirer des colonnes sans porter atteinte à l'intégrité des données et/ou des applications les exploitant. Il existe des frameworks ou des API, dans différents langages, permettant de palier à ces soucis (Hibernate [11]) mais ceux-ci sont trop souvent lourds et demandent tout de même, au final, une maintenance coûteuse.
Pour des raisons essentiellement techniques et afin de faciliter les travaux de maintenance des applications, de réutilisabilité des codes et de souplesse de développement, l'équipe technique de l'équipe CISMeF a décidé de refondre son système d'information en s'inspirant largement des outils du web sémantique comme la logique de description, les ontologies et ses formats tels que OWL 2.8.4, les triplets RDF 2.8.2, etc.
En effet, l'expérience acquise depuis 2005 dans le domaine des SOC, puis à partir de 2009, dans le domaine des ontologies puis via ces travaux de thèse a permis de concevoir une vue abstraite de toutes les données à gérer. Cette vision reste très proche de la programmation orientée objet, des ontologies et de la vision 3M développée dans ce mémoire. La méta-modélisation des SOC a ouvert une perspective de généricité dans le stockage, la gestion et la RI de données hétérogènes au sein d'un même système.

La gestion de données via un modèle unique est un défi important mais qui possède plusieurs avantages :

- Le nombre d'entités (donc de tables) est très faible et invariant ;
- Le nombre d'attributs de ces entités est également très faible et invariant ;
- L'ajout d'entité ou d'attributs ne perturbent en rien les modèles conceptuel, logique et physique de la base de données et, si les applications ont bien été développées, les modifications en découlant, sont très simples, voire nulles ;
- La recherche d'information est facilitée par la création de programmes spécia-

11. Jboss Community, http ://www.hibernate.org/, août 2013

lisés et valables pour toutes les données, quelque soit leur domaine d'application ;
- Le système d'information est générique ce qui permet une conception, une administration et un développement plus aisés des applications exploitant les données ;

Il présente aussi quelques contraintes importantes :
- Il y a peu de tables mais donc beaucoup de lignes ce qui implique de bons moyens techniques (serveurs adaptés et SGBD performant) ;
- La complexité des modèles se reporte au niveau du typage des données (chaque ligne - ou enregistrement - doit être typée pour savoir de quelle entité il s'agit).

3.6.1 Description du modèle générique de données

Deux parties principales composent ce modèle (Figure 3.5) :
- la partie MODEL qui décrit les modèles conceptuels, les méta-données mais aussi des règles et des contraintes (équivalent des classes des ontologies ou de la POO ou de la T-Box (T) dans la logique de description) ;
- la partie OBJECT qui stocke les données (équivalent des instances des ontologies ou de la POO ou de la A-Box (A) dans la logique de description).

Quatre tables décrivent les modèles conceptuels et les « entités système » de la base de données : `TB_MODEL`, `TB_MODEL_DATATYPE_PROPERTY`, `TB_MODEL_OBJECT_PROPERTY` et `TB_MODEL_INHERITANCE`.
Cinq tables stockent les « vraies » données : `TB_OBJECT`, `TB_DATATYPE_PROPERTY`, `TB_OBJECT_PROPERTY`, `TB_HIERARCHY`, `TB_INDEXING`.
Trois autres tables servent à la gestion du système ou aux applications tierces (cache, logs, ...) : `SYS_CACHE_UPD_QUEUE`, `SYS_LOGGER` et `SYS_PARTITIONS`. Trois vues permettent la gestion d'authentification des utilisateurs sur les applications clientes : `V_ROLES`, `V_USERS` et `V_USER_ROLES`. Ces tables et vues n'apparaissent pas dans le modèle physique de données ci-dessous.
On peut classer les neuf tables principales en trois catégories :
- Les tables créant les références (classes et instances) : `TB_MODEL` et `TB_OBJECT`
- Les tables contenant les attributs (chaines de caractères, entiers, dates, booléens, etc.) : `TB_MODEL_DATATYPE_PROPERTY` (MDP) et `TB_DATATYPE_PROPERTY` (DP)
- Les tables contenant les relations entre références : `TB_MODEL_OBJECT_PROPERTY` (MOP), `TB_MODEL_INHERITANCE` (MIN), `TB_OBJECT_PROPERTY` (OP), `TB_HIERARCHY` (HIE) et `TB_INDEXING` (IND)

Deux éléments principaux servent à identifier les « classes » et les « instances » de la base :
- les TYPE_ID (TI) qui sont les identifiants uniques des types définis dans la partie MODEL ;

– Les RDF_RESOURCE qui sont les identifiants uniques des objets définis dans la partie OBJECT.

Partie MODEL

La partie MODEL se substitue à la méthodologie classique où les objets sont transformés en entités avec un certain nombre d'attributs convertis en colonnes. Ici, chaque entité et chacun de ses attributs sont transformés en TYPE_ID (TI). Les TI sont des « types » de tous genres (types d'objets, types d'attributs, types de relations, etc.). Ainsi, chaque table du modèle physique contient une colonne TYPE_ID qui définit le type de la donnée. On définit le TYPE_DOMAIN (TD) d'un TI comme sa « portée », c'est-à-dire à quelle table il s'applique. Ainsi, un TI dont le TD sera DATATYPE_PROPERTY ne pourra être utilisé que dans la table `TB_DATATYPE_PROPERTY`.
Une des particularités de ce modèle physique repose sur sa capacité à intégrer n'importe quel modèle conceptuel mais surtout à en stocker plusieurs, indépendants ou non, dans le même schéma. Afin de faciliter la représentation de ces « sous-modèles », la notion de *domaine de base de données* a été créée (à ne pas confondre avec le type de domaine - TD - décrivant la portée d'un TI). Ainsi, le TI DOMAIN est le supertype de tous les domaines de la base. Chaque domaine regroupe un ou plusieurs TI ou sous-domaines (décrits ci-après) via un MOP de TI APPLICATION_DOMAIN. Les sous-domaines sont facultatifs et permettent d'avoir un niveau intermédiaire entre domaine et TI. Typiquement, les SOC possèdent tous un TI (TER_*) ayant pour TD MODEL et pour domaine DOMAIN_TER. Les TI utilisés réellement dans les données (types de descripteurs) sont reliés à leur sous-domaine via un MOP de TI BELONGS_TO.
Afin de décrire précisément un schéma de modèle conceptuel, plusieurs TI ont été créés en plus des notions de domaines et de sous-domaines. Premièrement, chaque entité à modéliser sera classiquement représentée par un TI de TD OBJECT. Ce TI sera relié soit à un sous-domaine, soit directement à un domaine. Ensuite, décrire les attributs et les relations éventuels s'appliquant aux TI créés peuvent se faire via les TI suivants :

– HAS_ATTRIBUTE : pour spécifier un TI (de TD DATATYPE_PROPERTY) d'attribut ;
– HAS_RELATION : pour spécifier un TI (de TD OBJECT_PROPERTY) de relation, de hiérarchie ou d'indexation.

Le tableau 3.1 montre un extrait de la table `TB_MODEL_OBJECT_PROPERTY` utilisant des triplets définissant les règles de représentation des domaines et des schémas conceptuels pour un SOC. Ici, il s'agit d'un type de concept du SOC « Human Phenotype Ontology » (HPO). La table `TB_MODEL_DATATYPE_PROPERTY` permet de décrire précisément les TI et donc, les méta-données. Il est ainsi possible de leur

TYPE_ID_SOURCE	TYPE_ID	TYPE_ID_TARGET
TER_HPO	APPLICATION_DOMAIN	DOMAIN_TER
T_DESC_HPOTERM	BELONGS_TO	TER_HPO
T_DESC_HPOTERM	HAS_ATTRIBUTE	RDFS_LABEL
T_DESC_HPOTERM	HAS_ATTRIBUTE	T_UF_HPO
T_DESC_HPOTERM	HAS_RELATION	T_REL_HPO_TO_FMA

TABLE 3.1 – Extrait de la table `TB_MODEL_OBJECT_PROPERTY` représentant une partie du modèle de la HPO

attribuer des libellés, et ce, dans plusieurs langues.
Afin de pouvoir représenter les TI dans la partie OBJECT, il est nécessaire d'instancier des objets dits « abstraits » dans `TB_OBJECT`. Il suffit de créer un nouvel objet qui aura comme RDF_RESOURCE le TI concaténé à « _OBJ » et qui aura pour TI le TI lui-même.

Partie OBJECT

C'est dans cette partie du modèle que sont stockées les données « réelles ». Chaque nouvel objet doit être référencé dans la table `TB_OBJECT` avec son TI pour le qualifier. La table la plus volumineuse est `TB_DATATYPE_PROPERTY` puisqu'elle contient tous les attributs des objets : termes, dates, hyperliens, textes, valeurs numériques, etc. La gestion du multi-linguisme tient uniquement au fait d'avoir autant de lignes dans DP qu'il y a de langues disponibles pour chaque terme. Pour cela, la colonne XML_LANG est utilisée ; la norme ISO 639-1 [12] (étendue à la norme ISO 639-3 pour les langues non officielles) permet de décrire les langues grâce à deux lettres (« fr » pour le français, « en » pour l'anglais, etc.) (voir le Tableau 3.2 pour l'exemple des libellés préférés du Descripteur MeSH « coeur » (D006321)). Cette table est aussi la plus complexe à gérer au point de vue technique (cf. 3.6.3). La structure de `TB_DATATYPE_PROPERTY` a été inspirée du modèle EAV (pour Entity-Attribute-Value), lui même très proche du triplet RDF. Le modèle EAV, décrit au début des années 1980 [Stead *et al.*, 1982], repose alors sur une Entité (ici RDF_RESOURCE), un Attribut (ici le TYPE_ID) et une Valeur (ici, la colonne VAL) et peu ainsi représenter n'importe quelle valeur d'attribut d'objet en la typant précisément via une table de correspondances. En rajoutant donc la colonne XML_LANG, ce modèle devient multi-lingue (EAVL pour Entity-Attribute-Value-Language). Par ailleurs, notre approche est très similaire à celle développée dans [Dinu & Nadkarni, 2007].
Les trois tables stockant les relations entre objets sont très proches également de la modélisation en triplets RDF avec une colonne pour désigner le premier objet en relation (RDF_RESOURCE_SOURCE), une deuxième colonne pour le second objet (RDF_RESOURCE_TARGET) et une colonne typant la relation entre ces deux objets (TYPE_ID). Bien qu'elles soient basées sur le même modèle, ces trois tables

12. http://www.loc.gov/standards/iso639-2/php/code_list.php

RDF_RESOURCE	TYPE_ID	XML_LANG	VAL
MSH_D_006321	RDFS_LABEL	cs	srdce
MSH_D_006321	RDFS_LABEL	de	Hertz
MSH_D_006321	RDFS_LABEL	en	heart
MSH_D_006321	RDFS_LABEL	es	corazón
MSH_D_006321	RDFS_LABEL	fi	sydän
MSH_D_006321	RDFS_LABEL	fr	coeur
MSH_D_006321	RDFS_LABEL	hr	srce
MSH_D_006321	RDFS_LABEL	it	Cuore
MSH_D_006321	RDFS_LABEL	lv	Sirds
MSH_D_006321	RDFS_LABEL	nl	Hart
MSH_D_006321	RDFS_LABEL	pt	coração
MSH_D_006321	RDFS_LABEL	se	Hjärta

TABLE 3.2 – Extrait de la table `TB_DATATYPE_PROPERTY` représentant les libellés préférés en plusieurs langues du Descripteur MeSH « coeur » (D006321)

n'ont pas été rassemblées en une seule table pour éviter des confusions, améliorer les performances et faciliter l'administration. Par ailleurs, chaque table de relations possède une petite particularité. Pour `TB_OBJECT_PROPERTY`, il y a deux colonnes supplémentaires pour la pré- et post-coordination (spécifique au codage terminologique donc). Pour `TB_HIERARCHY`, il y a une colonne spéciale pour stocker les chemins hiérarchiques (cf. 4.7.2). Pour `TB_INDEXING`, il y a deux colonnes pour la post-coordination et une colonne pour le poids (ou score) de l'indexation.

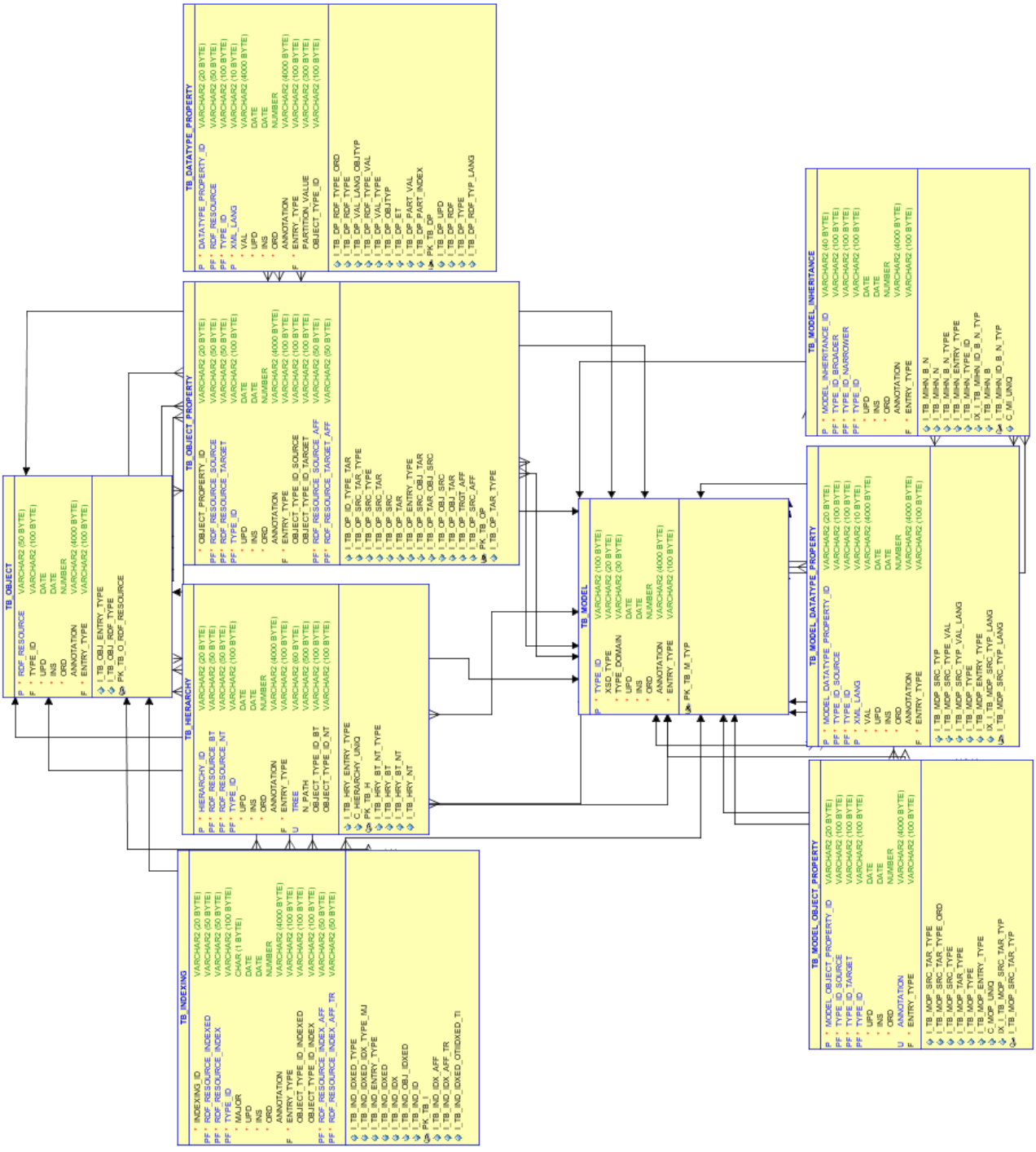

FIGURE 3.5 – Modèle Physique de Données du SI de CISMeF

3.6.2 Relations *n-n*

L'entité `TB_OBJECT_PROPERTY` permet de stocker des relations *1-1* et *1-n*. Cependant, pour les rares cas de relations *n-n*, il a fallu trouver une solution car un triplet seul ne peut modéliser cela. Il existe deux solutions possibles : (i) créer un concept factice correspondant à au moins l'une des unions des concepts à relier, (ii) ou utiliser un quatrième élément pour compléter le triplet (qui devient donc quadruplé). Nous avons opté pour cette seconde solution car moins complexe à mettre en œuvre. Les concepts de chaque « côté » d'une relation *n-n* sont rassemblés via un identifiant commun `ID_OBJECT_PROPERTY`. Les couches métiers peuvent ensuite interpréter cela pour regrouper les concepts.

3.6.3 Considérations techniques

Le choix d'implémentation de ce modèle logique ne s'est pas fait à la légère mais a été guidé par l'expérience acquise par l'équipe CISMeF mais également par des tests, grandeurs natures, et par de la veille bibliographique et technique sur le sujet. Stocker l'information de façon générique est pratique mais implique un coût élevé au développement initial mais aussi souvent en terme de matériels informatiques. Les systèmes de demain sont aujourd'hui en cours de conception et constituent un défi majeur. En matière de stockage d'information, d'édition et de RI, plusieurs approches sont à notre portée.
Les tests effectués au sein de l'équipe CISMeF révélaient encore récemment que la base de données relationnelle, bien administrée et maîtrisée, restait encore aujourd'hui l'alternative technique la plus sûre et la plus performante. Les choix d'implémentation du modèle logique présenté dans ce mémoire ont été guidé par l'expertise poussée des ingénieurs de l'équipe au SGBD Oracle avec ses options particulièrement bien adaptées à la RI.
Cependant, comme les technologies évoluent extrêmement vite sur ce sujet, de nouvelles études sont en cours pour déterminer les performances et les coûts possibles afin de migrer et d'exploiter les données du SI du CISMeF sur des outils comme Infinispan (cache distribué), SPARQL (entrepôts RDF) ou encore des SGBD NoSQL. L'objectif est de trouver un éventuel remplaçant à Oracle qui, malgré ses performances, souffre de plusieurs désavantages : coût très onéreux de la licence et quelques défauts techniques majeurs (champs textes limités à 4000 caractères, partitionnement dynamique obscur, ...).

Comparaison avec d'autres technologies

Infinispan[13] est un système de cache distribué permettant la mise à disposition rapide des données. Utilisé pour l'instant seulement comme couche de cache dans le

13. http://infinispan.org/

SI du CISMeF, il peut également servir d'outil de RI via un module dédié utilisant Hibernate Search [14] (index) et le moteur Lucene [15] (RI proprement dite).
Le langage SPARQL (pour *Protocol and RDF Query Language*) permet d'effectuer des requêtes sur un entrepôt de données RDF via un protocole spécifique. Les triplets RDF constituent aujourd'hui une structuration de la données particulièrement utilisée dans le monde. Ils permettent de tout représenter sous une forme textuelle facilement échangeable. Il est donc tout à fait naturel qu'un langage ait été développé et adopté comme standard pour exploiter ces données. Trois recommandations ont donc été conçues autour du langage d'interrogation, du format XML de résultat et du protocole utilisé [16].
Les bases de données NoSQL sont des entrepôts non relationnels et n'utilisent (quasiment) pas de SQL [17]. Les données y sont stockées sous forme de tables de hachage (associations clé-valeur) reparties, le plus souvent, entre plusieurs machines en réseau, cela permettant la réplication et la mise à disposition rapide des informations (cache). Cependant, cette technologie possède quelques désavantages comme une moins bonne fiabilité aux transactions ou encore la difficulté à écrire certaines requêtes complexes [Leavitt, 2010]. Par ailleurs, il existe un grand nombre de modèles de données NoSQL en fonction des besoins (orientés colonnes, orientés graphes, orientés documents, etc.) et cela rend les comparaisons difficiles. Citons quelques SGBD NoSQL parmi les plus connus et les plus utilisés : CASSANDRA [18], MongoDB [19] ou encore SimpleDB [20].

Ces systèmes et technologies ont montré de très bons résultats, surtout pour la partie NoSQL, extrêmement utilisée dans des applications gérant du Big Data, comme Facebook, Amazon ou Twitter. Cependant, cela demande des matériels onéreux et en grand nombre puisque ce système fonctionne souvent en *cloud computing*. De plus, ces applications sont très réactives mais ne correspondent pas à de la RI complexe, basée sur des aspects sémantiques, des vocabulaires contrôlés et des méta-données nombreuses et spécifiques. Des comparaisons poussées sont en cours de réalisation par des étudiants ingénieurs au sein de l'équipe CISMeF.

14. http://hibernate.org/search/
15. http://lucene.apache.org/
16. http://travesia.mcu.es/portalnb/jspui/handle/10421/7464
17. Le nom NoSQL peut paraître étrange puisqu'il s'agit en fait de la contraction de *Not Only SQL*
18. http://cassandra.apache.org/
19. http://www.mongodb.org/
20. https://aws.amazon.com/fr/simpledb/

Volumes

Le tableau 3.3 récapitule le nombre de lignes par tables (au total et uniquement sur la partie S3M) du modèle générique de données ainsi que, à titre informatif, le nombre de TI utilisés dans ces tables. On peut voir que le nombre de lignes dans `TB_DATATYPE_PROPERTY` avoisine les 20 millions. Comme exposé plus tard dans ce mémoire, la majeure partie de ces lignes concerne les termes et autres attributs des concepts de SOC. Par ailleurs, on peut remarquer que la majeure partie de la BDD est peuplée par les données des SOC (72,3% des objets et la quasi totalité des relations) ; les nombres de 100% pour les tables `TB_HIERARCHY` et `TB_INDEXING` sont facilement explicables par le fait que seuls les SOC sont hiérarchisés et que l'indexation se fait sur les documents à l'aide des concepts terminologiques.
Déplacer la complexité de la représentation d'entités et d'attributs en lignes, au lieu de tables et colonnes, pose donc inévitablement des problèmes de volumétrie. Cela est d'autant plus important que le SGBD choisi est relationnel et que les applications l'exploitant sont avant tout des moteurs de recherches à haute performance. La RI demande des technologies et des méthodes complexes pour obtenir des résultats pertinents dans un temps minimal pour l'utilisateur final. D'ailleurs, l'un des objectifs de ces travaux de thèse était celui du passage à l'échelle entre l'exploitation d'un serveur intégrant 5 terminologies en français/anglais (InterSTIS) vers un serveur gérant plus de 50 terminologies dans plus de 20 langues. Il pourrait donc sembler étonnant d'avoir fait le choix d'un modèle plus compact, donc plus coûteux en nombre de lignes, pour pouvoir stocker autant d'informations.
La réponse à ce problème technique repose sur les technologies des index et de partitionnements proposées notamment par Oracle.

Table	Nb de TI	Nb de lignes total	Nb de lignes S3M
TB_OBJECT	306	3 037 741	2 195 756 (72,3%)
TB_DP	400	19 885 879	12 961 958 (65,2%)
TB_OP	317	7 645 806	6 764 191 (88,5%)
TB_HIERARCHY	48	1 615 438	1 615 438 (100%)
TB_INDEXING	3	7 478 338	7 478 338 (100%)
TB_MODEL	1 895	1 895	1 060 (55,9%)
TB_MODEL_DP	64	9 966	5 710 (57,3%)
TB_MODEL_OP	43	7 793	5 999 (77%)
TB_MODEL_INHERITANCE	1	1 035	1 025 (99%)

TABLE 3.3 – Nombres de TYPE_ID et de lignes des neuf tables principales du modèle physique de données (environnement de production CISMeF au 27 mai 2014).

Index de domaine

En plus des index « classiques », Oracle propose des « index de domaine » permettant des recherches très rapides. Il existe plusieurs « domaines » : spatial, d'image

ou encore textuel. Les index créés prennent donc en considération la nature, la structure et la valeur des données stockées. Dans notre modèle, il s'agit d'un index de domaine textuel (appelé communément *Oracle Text*) sur la colonne `VAL` de la table `TB_DATATYPE_PROPERTY`. Pour aller plus loin, il s'agit en fait de plusieurs index de domaines locaux partitionnés.

Partitionnement

Dans l'univers des SGBD, une partition est une division logique d'une table d'une BDD. Les différentes partitions sont, le plus souvent, indépendantes et permettent la gestion automatique de plusieurs tables logiques distinctes (voir la Figure 3.6 illustrant la différence entre table non-partitionnée et table partitionnée). Cela possède plusieurs avantages : s'abstraire des limites de stockage des *tablespaces* et gagner en performances via le parallélisme et la présélection de partitions à requêter. Il existe trois grands types de partitionnements : (i) par tranche ou gamme (*range*), (ii) par liste et (iii) par hachage (*hash*). Les index créés sur des tables partitionnées peuvent être globaux (index classiques sur la table entière) ou locaux (index spécifiques à chaque partition).
Le partitionnement, les index locaux partitionnés et les index de domaines locaux partitionnés permettent donc de créer des « sous-tables » logiques invisibles pour les développeurs et assurent de très bonnes performances, même sur des tables gigantesques.
Toute la complexité et la volumétrie de la table `TB_DATATYPE_PROPERTY` de notre

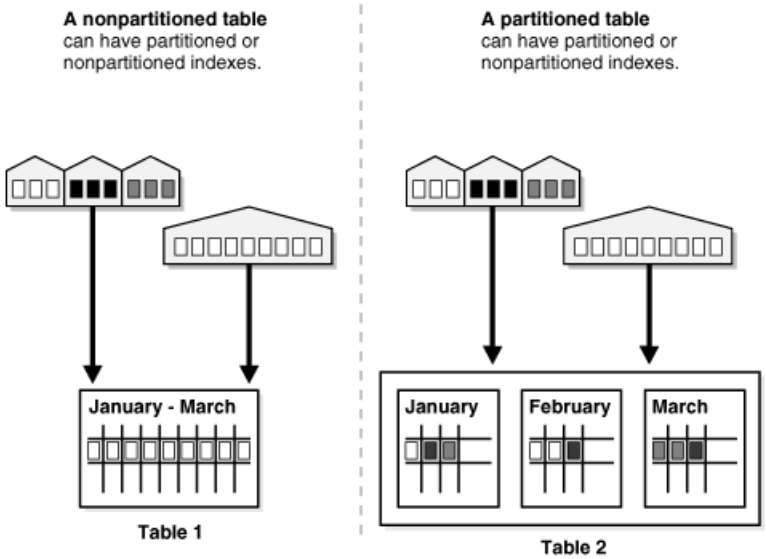

FIGURE 3.6 – Les tables partitionnées ou non peuvent avoir des index partitionnés ou non *(illustration extraite de la documentation Oracle en ligne)*

modèle logique sont alors « compensées » via le mécanisme de partitionnement et d'index locaux partitionnées. Concrètement, `TB_DATATYPE_PROPERTY` est parti-

tionnée par *range*, chaque tranche correspondant à l'union des colonnes suivantes : `TYPE_ID`, `OBJECT_TYPE_ID` et `XML_LANG`. Ainsi, une colonne composite (`PARTITION_VALUE`) stocke automatiquement la concaténation de ces trois colonnes et chaque nouvelle ligne insérée est rangée automatiquement dans la partition correspondante. Ce mécanisme est particulièrement bien adapté au contexte du S3M. En effet, il existe un grand nombre de `PARTITION_VALUE` différentes puisqu'il y a beaucoup de types de concepts, de types d'attributs et de langues intégrés. Chaque nouvelle combinaison entraîne la création d'une nouvelle partition et donc d'un sous-ensemble de données. Les requêtes SQL impliquant `TB_DATATYPE_PROPERTY` gagnent alors en performance lorsqu'il est possible de savoir sur quelle(s) partition(s) il faut chercher. Il existe, à ce jour, 2 093 `PARTITION_VALUE` différentes dans `TB_DATATYPE_PROPERTY` dont 1 910 relatives aux SOC du S3M (91,2%). Enfin, il est à noter que le système des partitions sous Oracle n'est pas nativement dynamique. Cela a été un problème technique complexe à résoudre. Nous avons créé une surcouche applicative à l'aide d'une table dédiée, de scripts PL/SQL [21] et d'un système de déclencheurs.

3.6.4 Modèle générique, outils génériques

L'un des gros avantages d'un modèle générique est le développement d'outils génériques l'exploitant. Il est nécessaire de concevoir des codes informatiques relativement abstraits, puisque génériques, mais donc bien moins volumineux. Leur administration (correction d'anomalies, évolutions, modularité, optimisation) est grandement facilitée.
Ainsi, lorsque l'équipe technique de CISMeF a débuté l'implantation du modèle physique générique présenté ici, la première étape a été de concevoir et développer une API (*Application Programming Interface*) en Java permettant de se connecter, de manipuler et de faire des requêtes sur les entités de la BDD. Cette brique logicielle centrale est appelée « DBCore ». Cette API est donc utilisée par la totalité des applications gravitant autour du S3M.

3.6.5 Utilisations du modèle générique de données

L'utilisation et l'impact de ce modèle physique de données sur le SI de CISMeF se sont avérés plus importants que prévus. En effet, la généricité des outils couplés à la base de données, la souplesse offerte par la partie MODEL, la gestion native du multi-linguisme et les très bonnes performances ont fait de ce modèle un outil indispensable au développement de toute nouvelle application. Ainsi, le SI entier du

21. *Procedural Language/Structured Query Language*, langage utilisé pour construire des procédures et fonctions sous Oracle, soit des petits programmes séquentiels pouvant contenir du SQL

CISMeF repose sur l'implantation de ce modèle, qu'il s'agisse du S3M, de la base de données documentaires ou encore du domaine des Dossiers Patients Informatisés (DPI).
Voici les différents domaines exploités via ce modèle, tous environnements confondus :

- Terminologies/Ontologies (S3M) ;
- CISMeF (catalogue et données associées comme les éditeurs, etc.) ;
- Presse Médicale (outil d'évaluation) ;
- Dossier Patient Informatisé ;
- BDBfr, base bibliographique en français (cf. 5.4.3) ;
- Évaluations (utilisé pour évaluer différentes applications) ;
- Domaine « Commun » : autres entités communes à beaucoup d'applications (langues, pays, etc.) ;
- Administration système (utilisateurs, libellés d'applications, etc.).

La quasi-totalité de ces domaines est associée au S3M pour l'indexation ou le codage et bénéficie alors de toutes les fonctions et contenus des SOC.

3.7 Synthèse du chapitre

Dans ce chapitre, plusieurs systèmes 3M ont été présentés, tous issus du domaine de la Santé. Nous avons ensuite vu que l'une des stratégies pour gérer la multi-terminologie consistait à créer un méta-modèle. Dans notre approche, nous proposons un méta-modèle 3M compact et flexible. Celui-ci a été implémenté dans un modèle logique de données générique novateur. Ce schéma est également très compact et flexible puisque notamment inspiré par le méta-modèle 3M. Son implémentation sous le SGBD Oracle a été un succès et offre de nombreuses possibilités quant à la gestion des informations stockées.
Suite à l'élaboration de ces modèles et de l'étude des systèmes 3M proches, la prochaine étape consistait à intégrer des SOC dans ce S3M. Pour cela, il a fallu concevoir une méthodologie d'intégration reproductible de SOC en SOC et permettant de gérer le plus finement possible les étapes de mises à jour et d'enrichissement des contenus.

Chapitre 4

Intégration des SOC

Sommaire

Dans ce chapitre, il s'agit de présenter les méthodes et les outils développés et utilisés pour intégrer les SOC dans le S3M.
Les choix des SOC à intégrer ont été guidés par différents critères : utilité dans les domaines d'études de CISMeF (indexation, codage du Dossier Patient Informatisé, etc.) et dans ses projets de recherche, utilisation par la communauté et disponibilité sous forme électronique de la version la plus récente.
Les différents SOC choisis ont été intégrés dans le S3M. La liste complète est disponible dans le Chapitre des résultats (cf. 6.4.1).
Le S3M est de fait le cœur des réalisations de ces travaux et constitue aujourd'hui la partie centrale du SI du CISMeF dont la plupart des applications se basent sur des vocabulaires contrôlés. Les parties ci-après décrivent les différentes étapes et détaillent les outils nécessaires à une intégration ou une mise à jour d'un SOC dans le S3M.

4.1 Choix des SOC à intégrer

Intégrer un nouveau SOC dans le S3M ne se fait pas au hasard, ni par pur souhait de faire grossir la taille du serveur. En effet, chaque SOC est intégré pour un but précis. Ils sont classé en 4 catégories (les SOC pouvant bien sûr appartenir à plusieurs catégories à la fois) :

- les SOC indispensables, car utilisés dans le SI du CISMeF pour l'indexation de documents (MeSH, ATC, etc.) ;
- les SOC de référence, car utilisés dans bon nombre de systèmes, soit en France, soit dans le monde (CIM-10, CIM-9, CCAM, LOINC, etc.) ;
- les SOC utiles pour des projets de recherche (NCIT, OMIM, terminologies d'interface, etc.) ;
- les SOC intéressants du point de vue enrichissement de la connaissance, travaux de recherche (HPO, HRDO, SNOMED CT, etc.).

Le S3M n'a donc pas pour vocation de s'ouvrir à la communauté afin de permettre l'intégration de nouveaux SOC comme peut le proposer BioPortal. Il ne s'agit pas non plus d'un catalogue de SOC mais bien d'un serveur permettant leur utilisation dans des buts bien précis, et ce, en insistant sur leur qualité. En effet, comme présenté ici, chaque nouveau SOC intégré est analysé puis contrôlé pour enfin être enrichi afin d'offrir aux utilisateurs le meilleur service possible.
Par ailleurs, il est important de noter que l'ajout de nouveau SOC s'est fait au fil de l'eau pour ces travaux de thèse ; cela a été un problème au point de vue gestion du

temps et pour l'adaptabilité des méthodes et outils développés mais cela s'est aussi avéré un avantage car la méthodologie d'intégration a été éprouvée un bon nombre de fois, ce qui participe, de fait, à sa validation.

Les droits sur les SOC

La plupart des SOC ont une valeur intrinsèque car collecter et structurer la connaissance dans un domaine précis demande beaucoup de travail d'expertise. Cette valeur est intellectuelle mais également financière et constitue un enjeu majeur pour certains éditeurs, publics ou privés. Même si la plupart des éditeurs publics nationaux ou internationaux mettent à disposition gratuitement leurs SOC, leur utilisation est souvent régie par des accords ou des licences libres. D'autres SOC, en revanche, sont clairement protégés par des licences très contraignantes et nécessitant des contrats écrits et éventuellement des transactions financières. Ces montants peuvent être parfois importants dans le cas d'entreprises privées (MedDRA, SNOMED CT, ...). Il s'avère que, à des fins de recherche académique, il est souvent facile de pouvoir exploiter gratuitement les SOC. Il s'agit cependant d'être conscient de l'impact de ce genre d'accord. Ainsi, dans le cadre de ces travaux de thèse, à chaque nouvelle intégration de SOC, la question des droits d'accession et d'exploitation s'est posée. Afin de protéger au maximum des SOC « sensibles » ainsi que le contenu ajouté par les experts de l'équipe CISMeF, l'une des contraintes du S3M a été de concevoir un système de protection des données et d'authentification aux applications les présentant.

4.2 Intégration des terminologies et ontologies (ETL)

Dans la méthodologie mise en place lors de ces travaux, l'intégration d'un SOC dans le SI s'effectue en six étapes successives au maximum. Elles constituent ce que l'on appelle communément un ETL pour « Extraction - Transformation - Loading ». Il s'agit en fait de 3 étapes principales autour desquelles des étapes supplémentaires ou optionnelles ont été ajoutées :

1. Constitution du modèle conceptuel du SOC ;
2. Formatage des données (optionnelle) ;
3. Extraction et transformation des données ;
4. Contrôles des données et post-traitements ;
5. Intégration proprement dite dans le S3M ;
6. Enrichissement du SOC (optionnelle) ;
7. Pré-calculs pour les applications (batchs).

La Figure 4.1 illustre les étapes 1 à 5 ; il s'agit de créer un modèle terminologique héritant du méta-modèle 3M en se basant sur la documentation, les données ainsi que

sur l'expertise de certains spécialistes du SOC à intégrer. Parallèlement à cela, il est souvent nécessaire de formater les données sources (encodage, format intermédiaire, etc.). Ensuite, un parseur Java est écrit spécifiquement pour chaque nouveau SOC pour créer un fichier unique contenant toutes les données extraites, organisées et formatées. Puis, un outil permet de lire ce fichier pour effectuer différents contrôles pour valider l'intégrité des données. Enfin, la dernière étape consiste en l'intégration dans la base de données du S3M du SOC via deux fichiers RDF/XML contenant les données et le fichier OWL contenant le modèle.
Toutes ces étapes sont détaillées précisément ci-après.

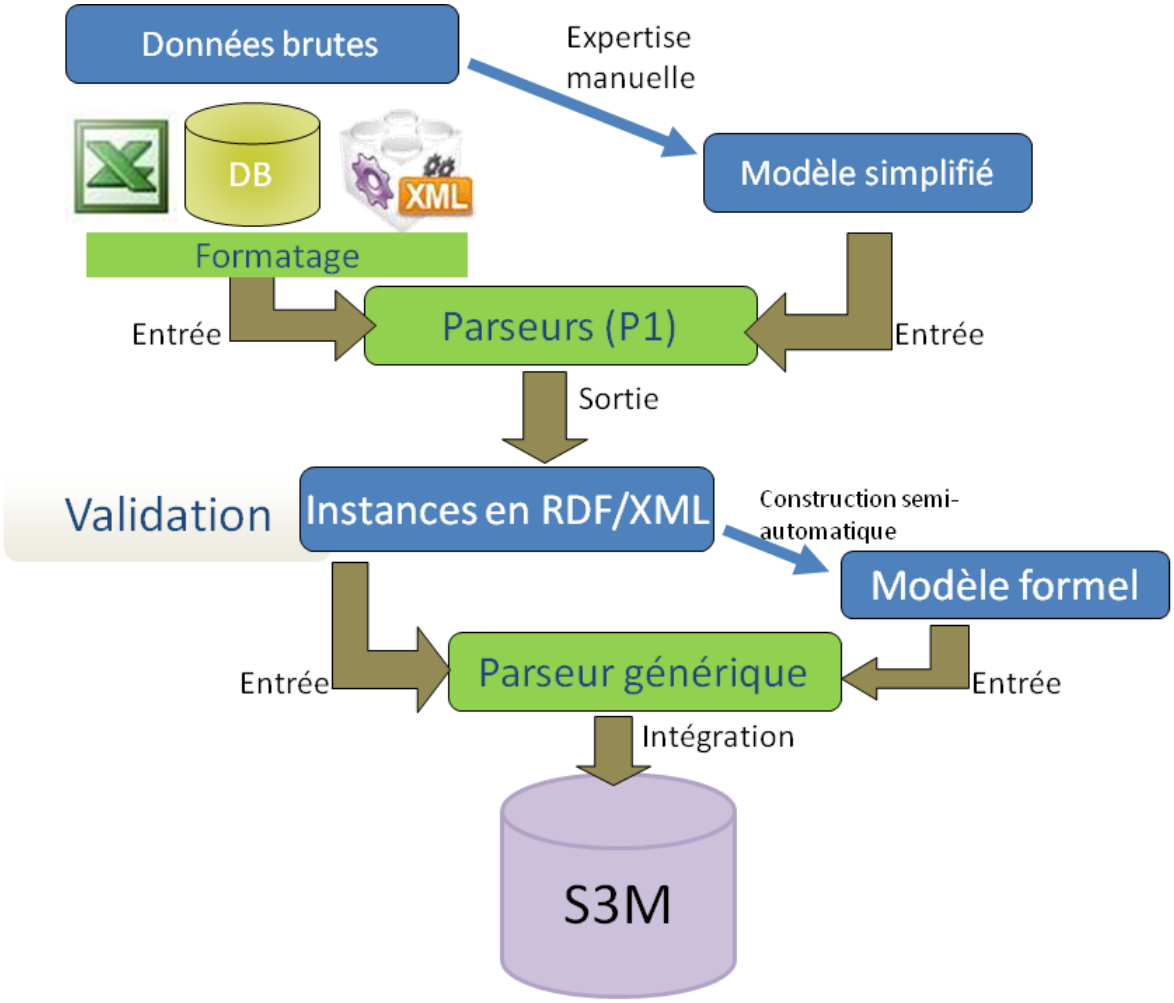

FIGURE 4.1 – Schéma de la méthodologie de modélisation, d'extraction des données de SOC puis d'intégration dans le S3M

4.3 Modèles spécifiques et méta-modèle 3M

Afin d'intégrer un nouveau SOC dans le système, la première étape consiste à appréhender le modèle spécifique de ce SOC. Cette étape est importante à deux titres ; premièrement, pour exploiter au maximum les fonctionnalités et les capacités d'un SOC et secondement, pour confronter ce modèle spécifique au méta-modèle 3M décrit avant. En effet, même si tous les modèles de SOC intégrés jusqu'à aujourd'hui sont compatibles avec ce méta-modèle, chaque nouveau SOC peut apporter un élément pouvant l'invalider ; le méta-modèle terminologique peut montrer ses limites à chaque nouvelle tentative d'intégration si, par exemple, un cas particulier n'a jamais été abordé.

Il n'y a pas de méthode définie pour appréhender un nouveau modèle de SOC. Cela est essentiellement dû au fait que chaque SOC est différent autant dans ses fonctions que dans sa structure et que les concepteurs du SOC n'ont pas forcément eu une démarche de modélisation lorsque le SOC a été créé. J'ai regroupé ici les différents types de modèles rencontrés jusqu'à aujourd'hui dans 4 grandes catégories ordonnées selon leur degré croissant de structuration :

- Modèles simplistes. Il s'agit de modèles n'ayant aucune réelle existence à l'origine mais que j'ai créés *de novo*. Ils contiennent généralement un type de concept qui correspond en fait à une entité dans la source originale des données. Exemples : les gènes dans la base de données NCBI, les phénotypes OMIM, les codes ADICAP, les différents types de IUPAC (constantes physiques, préfixes d'unités, quantités, termes et unités), etc. Ces entités correspondent très souvent à des éléments d'une banque de données, c'est-à-dire de données à plat sans aucune hiérarchie définie.
- Modèles hiérarchisés simples. Ce sont encore des modèles créés *de novo* mais ils correspondent à une structure légèrement plus complexe que les précédents puisqu'ils sont souvent organisés en hiérarchies (du type CHAPITRE - SOUS-CHAPITRE - DESCRIPTEUR par exemple). Ces modèles n'existent donc pas vraiment formellement mais sont assez simples à déduire. Par exemple, c'est le cas pour les SOC suivants : CIF, CIM-10, CCAM, MedDRA ou encore la SNOMED internationale 3.5 (voir l'exemple d'intégration en 4.8.2).
- Modèles terminologiques. Ce type de modèle est évidemment le plus facile à implémenter dans le méta-modèle 3M puisqu'il est nativement conçu pour être terminologique. Ces modèles définissent donc toujours leurs types de concepts de façon claire et bien souvent, aucune manipulation/modification n'est nécessaire pour les fondre dans le méta-modèle 3M.
- Modèles ontologiques. Paradoxalement, même si le modèle ontologique est bien souvent le plus complexe, le passage d'un modèle ontologique vers un modèle terminologique va simplifier les choses. En effet, comme décrit dans l'exemple de l'intégration de la FMA dans le S3M (4.8.3), si l'on voulait faire corres-

pondre un type de concept par classe ontologique, cela serait ingérable et incompatible avec une vision terminologique ; ainsi, on dégrade fortement le modèle ontologique en ne définissant alors qu'un seul (le plus souvent) type de concept pour ce SOC. Cela a été le cas pour la FMA, la SNOMED CT, le NCIT, HPO, la Gene Ontology ou encore RADLEX.

Le but de la modélisation spécifique consiste alors à comprendre la structure et les fonctions du SOC afin de déterminer :

- Les types de concepts
- Les types d'attributs de concepts
- Les types de relations et leurs directions (cf. 2.6)

Ces trois types d'éléments sont formalisés et stockés dans un fichier XML selon un schéma défini par nos soins. Aucun standard (non ontologique) n'existe à notre connaissance pour représenter un modèle terminologique de façon exhaustive ; le SKOS (cf. 2.8.3), par exemple, présente des éléments mais insuffisants pour nos besoins (pas de gestion de la polyhiérarchie contextuelle, pas d'attributs spécifiques natifs, etc.).

Toute la difficulté de cette étape sera en fait de comprendre le SOC par l'utilisation d'une documentation, de l'avis d'experts ou, à défaut, en analysant les données du SOC.

Étant donné que le méta-modèle 3M est terminologique, il faut, pour les modèles dits « simplistes » ou « simples », les « promouvoir » en modèles terminologiques. Pour les ontologies, on les « dégradera ».

4.4 Données

Une fois le modèle d'un SOC analysé et formalisé, il est nécessaire d'effectuer le traitement des données via différentes phases successives.

4.4.1 Sources de données

L'édition d'un SOC, et donc sa publication, dépend exclusivement de ses auteurs, de leur métier (documentaliste, médecin, biologiste, ...) et des objectifs du SOC. Même s'il existe des sources de données homogènes en terme de modélisation et de formats (UMLS en est le meilleur exemple), ce n'est pas le cas de la majorité des SOC. En effet, la plupart sont éditées dans des formats allant de formats ontologiques (OWL, OBO, ...) en passant par des bases de données relationnelles (MySQL, Oracle, Microsoft Access, etc.), aux fichiers Microsoft Excel, XML voir à des documents Microsoft Word ou Portable Document Format (PDF).

Il est important d'ajouter que les modèles des SOC sont rarement explicités dans la littérature ou dans un quelconque document et que cette partie d'expertise doit se faire spécifiquement à chaque intégration de nouveau SOC.

Ce pré-requis de disponibilité et de qualité d'une version électronique complète d'un SOC a été un problème, et ce, plusieurs fois au cours de ces travaux ; à chaque nouvelle demande d'intégration de SOC, la première question est toujours la même : quelle est la source de données et peut-on s'y fier ?
Un autre problème récurrent à la source de données est celui d'éventuels changements de formats entre deux versions. Comme expliqué plus bas, un parseur est écrit pour chaque source de données et si elle vient à changer lors d'une éventuelle mise à jour, le parseur est à adapter voire, à réécrire complètement.
Si, par ailleurs, le modèle lui-même vient à changer (ce qui est heureusement très rare), il faudra refaire l'étape d'intégration depuis l'étape 1.

4.4.2 Formatage des données

Pour beaucoup de formats de données des SOC, il est nécessaire d'effectuer une étape de formatage des données et ce, pour diverses raisons.
L'encodage des caractères peut poser des problèmes car suivant les langues et les pays, l'encodage varie. Or, dans un système multi-terminologique et multi-lingue, il est nécessaire d'utiliser un seul encodage. En l'occurrence, UTF-8 [1] a été choisi car il gère correctement tous les caractères nécessaires.
Certains formats natifs de SOC, typiquement Microsoft Word, PDF ou autres formats de texte plein, ne permettent pas d'extraire facilement les données et nécessitent donc la transcription dans un fichier plus structuré type tableau ou base de données. C'est souvent une opération fastidieuse nécessitant parfois des interventions manuelles sur les fichiers.

4.4.3 Extraction des données (P1)

Il s'agit de l'étape la plus complexe et la plus fastidieuse que j'ai nommée P1 (Parseur 1). Étant donné que les formats de données sont hétérogènes et ne respectent que très rarement les standards, extraire ces données reste une étape spécifique. Ainsi, pour chaque SOC, un « parseur » est écrit. Un parseur est un programme qui parcoure des données (fichiers, bases de données, etc.), le plus souvent ligne par ligne afin d'y effectuer un traitement quelconque.
Trois principaux problèmes peuvent se poser lors de l'élaboration d'un parseur 1 de SOC :

- La complexité de la structure des données en entrée. En effet, il est parfois difficile d'appréhender la structure initiale d'un SOC en fonction du format, surtout si elle n'est pas documentée. Par exemple, connaître les cardinalités des relations ou encore la nature de telle ou telle méta-donnée. Bien souvent, si

1. *UCS transformation format 8 bits*, http://www.utf-8.com/, norme ISO-10646 de gestion universelle de caractères des langues vivantes

ce type de problème se présente, il est nécessaire d'avoir recours à l'aide d'un expert du SOC ou de l'avis des principaux utilisateurs de ce type de SOC.
- La complexité mémoire, au sens informatique. Certains SOC sont très volumineux (SNOMED CT, MeSH, etc.) et peuvent donc constituer un défi de taille quant à la gestion de la mémoire du parseur. Il est donc nécessaire d'apporter une attention particulière aux algorithmes du parseur afin d'éviter des dépassements d'utilisation de la mémoire.
- La complexité en temps, au sens informatique. Généralement associée à la complexité mémoire, les SOC les plus volumineux sont longs à parcourir par le programme et très difficilement parallélisables. Quelques optimisations au niveau du code informatique peuvent encore aider mais la plupart du temps, lorsque le traitement nécessite plusieurs heures sur une machine locale « classique », il faut bénéficier de la puissance de certains serveurs du parc informatique de l'équipe CISMeF (cf. Annexe B.1).

Format de données intermédiaire

Dans le cadre du projet InterSTIS, un standard a été défini pour tous les partenaires du projet et ce, afin que chacun puisse transformer la source initiale d'un SOC en un fichier normé et compréhensible par les autres partenaires. Le choix s'est porté assez naturellement vers une modélisation en OWL « light » dans le format RDF/XML. Les modèles terminologiques sont alors représentés en OWL dans une partie du fichier et les différentes instances des classes sont représentés par des blocs RDF avec des espaces de noms (« namespaces »), et donc aussi des balises (« tags ») propres au projet InterSTIS.
Ce format étant suffisant et permettant de représenter au mieux les modèles et les données terminologiques, il a été conservé (avec quelques modifications mineures) dans la méthodologie d'intégration/mise à jour des SOC dans le S3M. Pour créer facilement ce genre de fichier de sortie, les classes Java du P1 héritent toutes d'une classe-mère implémentant des méthodes permettant d'écrire directement les concepts et leurs relations via l'API Java Jena [2]. Quelques exemples de sorties sont donnés dans la Section 4.8. L'utilisation d'un fichier intermédiaire, avant l'insertion réelle dans la base de données du S3M permet également d'effectuer une série de vérifications sur l'intégrité des données extraites (voire d'autres types de traitements).

4.4.4 Contrôle des données et post-traitements

Comme présenté dans le schéma d'intégration, une étape de contrôle des données (validation) est indispensable. Il a donc fallu concevoir et développer un programme local (en Java) appelé « Outil SMTS » (historiquement pour « Serveur

2. http://jena.apache.org/

Multi-Terminologique de Santé ») pour lire le fichier de sortie issue de P1 et proposer un certains nombre de post-traitements. L'idée de départ était d'utiliser un programme existant comme Protégé [3] mais cela s'est vite avéré impossible pour deux raisons principales : la taille des fichiers RDF/XML et donc la complexité mémoire nécessitait une configuration de machine beaucoup trop importante (par exemple, un fichier RDF/XML de 100 Mo ne pouvait être lu par Protégé à moins de 2 Go alloués à la JVM). Cela s'explique facilement car Protégé utilise toute la mémoire vive disponible pour charger l'ensemble de l'ontologie, ce qui n'a pas beaucoup d'intérêt pour nos besoins. La seconde raison pour laquelle nous avons choisi de développer un programme dédié, c'est la souplesse d'implémentation d'autres fonctionnalités à développer et la possibilité de réutiliser les classes Java dans d'autres applications. Voici les fonctionnalités importantes du programme :

- Parcourir un fichier de sortie d'extraction (P1) dans un espace mémoire restreint ;
- Analyser le fichier afin d'identifier les types de concepts, attributs et relations, ainsi que leurs métriques ;
- Permettre de séparer les concepts avec leurs attributs de leurs relations (pour P2) ;
- Permettre la création semi-automatique d'un fichier décrivant le modèle du SOC ;
- Permettre de détecter d'éventuelles incohérences dans le fichier (même concept défini plusieurs fois, concepts sans hiérarchies, etc.).

Avec ce programme, il est possible d'analyser un fichier de sortie de P1 en quelques secondes sur une machine locale et ainsi de contrôler l'intégrité des données. Une capture d'écran (Figure 4.2) présente l'analyse d'un fichier issu de P1 (ici la CISP-2, Classification Internationale des Soins Primaires, deuxième édition).

3. http://protege.stanford.edu

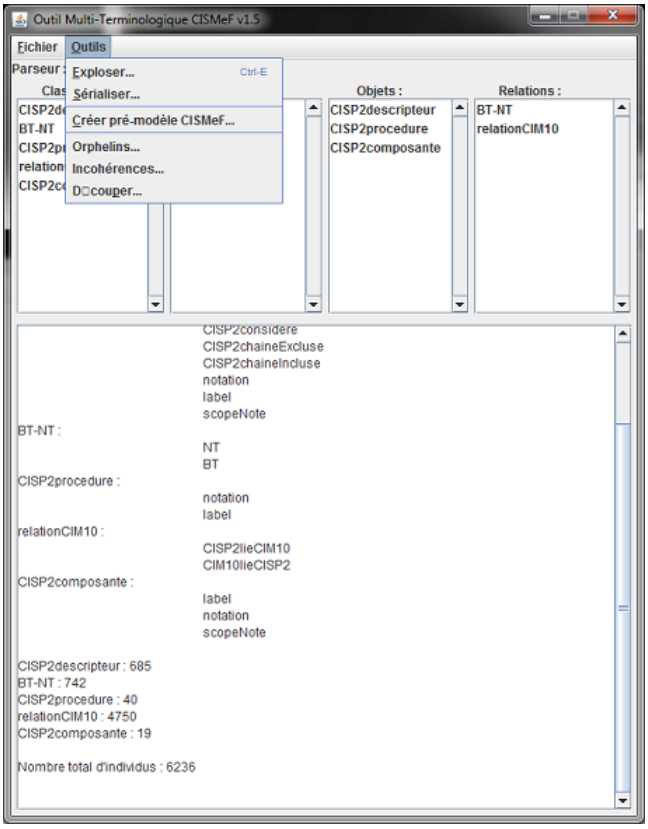

FIGURE 4.2 – Capture d'écran de l'Outils SMTS après chargement d'un fichier issu de P1.

4.4.5 Intégration dans la base de données du S3M (P2)

Une fois le contrôle des données effectué, l'étape finale d'intégration est celle du chargement des données dans la base de données (S3M). Le programme, appelé P2 pour Parseur 2, s'occupe de cette tâche. Il prend en entrée le fichier OWL/RDF généré par le P1 (et éventuellement modifié par l'étape de contrôle des données) et le fichier de modèle du SOC généré via l'Outil SMTS. Le chargement dans la base de données (ordres SQL INSERT, UPDATE, voire DELETE) s'effectuent en deux temps : insertion/mise à jour des « objets » puis insertion/mise à jour des « relations ». Ces deux étapes distinctes sont nécessaires pour assurer la contraintes d'intégrité des données dans le SGBD (contrainte de références des clés étrangères : il est impossible de créer une relation entre deux concepts s'ils ne sont pas déjà insérés dans la base de données). Le P2 est un parseur unique (générique) contrairement aux P1 spécifiques à chaque SOC. En effet, puisque le format des fichiers d'entrée est normé et que les méta-données sont décrites dans le fichier de modèle, le programme possède tous les éléments nécessaires pour insérer les données.
L'intégration de P2 dans le S3M génère des traces pour suivre en temps réel l'avancée des opérations. Il est alors possible de détecter d'éventuelles erreurs puis d'arrêter le programme en cours d'exécution afin de démarrer une phase de correction d'anomalie.
Cette étape d'intégration des données est régie par un certain nombre de règles pour gérer l'intégration, soit d'un nouveau SOC, soit d'une mise à jour de SOC. Ces règles sont détaillés ci-après (cf. 4.5.3) et constituent une partie de la méthodologie de versionnage conçue dans ces trauvaux.

4.5 Gestion des versions de SOC : versionnage

La gestion des versions de SOC ou versionnage (« versionning » en anglais) est une problématique récurrente dans le domaine de l'ingénierie des connaissances. Cela constitue un enjeu de taille puisque la grande majorité des SOC évolue. Cette évolution est normale et même nécessaire puisque la connaissance elle-même ne cesse d'évoluer et que les systèmes tendent à être modifiés, pour plus d'exhaustivité, de précision et de performance. Même si l'on peut considérer que la dernière version d'un SOC est la plus complète et donc, celle à utiliser, des contraintes techniques ou fonctionnelles imposent parfois d'utiliser des versions plus anciennes dans un SI donné. Il existe alors deux grands intérêts au versionnage de SOC :

- Comprendre et exploiter les différences entre deux versions de SOC (quels éléments ont changé ? éventuellement, pourquoi ? quels en sont les impacts ?) ;
- Stocker les différentes versions des SOC pour assurer le fonctionnement des systèmes basés sur une version donnée (rétro-compatibilité).

4.5.1 Définitions

Versionnage

Le versionnage se définit par la gestion des différentes versions d'une même ressource. Ces versions sont les différentes formes, ou niveaux d'évolution de cette ressource à des instants donnés. Il faut ici différencier les ressources en question, puisqu'il peut s'agir soit des SOC, des concepts terminologiques, voire même des attributs ou des relations de ces concepts.
Lorsque l'on parle de version de SOC, il s'agit de l'ensemble des concepts, de leurs propriétés et de leur représentation, valables à une date précise.

Historisation

L'historisation est le fait de stocker les renseignements nécessaires pour retrouver l'historique d'un élément donné. Dans le cas présent, il s'agit de tracer chaque modification (insertion, mise à jour, suppression) de chaque attribut ou relation d'un concept terminologique, et ce, en conservant la date et éventuellement l'ancienne valeur. Cela permet de garder un suivi extrêmement précis afin de comprendre pourquoi, quand et comment l'information a été mise à jour. L'intérêt principal réside aussi dans le fait qu'il est possible de revenir à une version *n-1*, *n-2* ou *n-m* d'une donnée.

4.5.2 Approches

Lorsque l'on parle de versionnage, on parle également d'édition de ressources. Les versions n'existent que par les modifications qu'apportent les éditeurs des SOC (éditeurs officiels ou tout autre expert gérant sa propre version d'un SOC). La façon dont sont gérées les versions est différente d'un SOC à l'autre. Certains sont mis à disposition régulièrement (une version « majeure » française du MeSH par an pour les traductions françaises par l'INSERM, une version par mois de la NABM, etc.). D'autres SOC sont très peu mis à jour et chaque nouvelle version engendre de gros changements, comme la Classification Internationale des Maladies (CIM) [4]. Ce SOC est assez particulier puisque la version officielle en France, par exemple, est la CIM-10 ($10^{\text{ème}}$ révision) alors qu'aux États-Unis, il s'agit encore de la CIM-9, qui évolue parallèlement. D'autres SOC sont mis à jour au fil de l'eau (c'est-à-dire constamment ou très régulièrement, comme la Gene Ontology par exemple) et d'autres SOC ne sont, quant à eux, plus du tout édités et ne subissent alors aucune nouvelle version (officielle, du moins). C'est le cas par exemple de la SNOMED internationale 3.5.
Des travaux ont été effectués, ou sont toujours en cours, pour concevoir des méthodes de détection des changements entre deux versions de SOC. Ces études sont

4. http://www.who.int/classifications/icd/en/

essentiellement axées sur les ontologies, plus faciles à manipuler de part leur formalisme. Il est tout à fait possible de calquer les méthodologies déjà développées dans ces travaux sur des terminologies. Ainsi, Klein distingue les changements qui affectent la conceptualisation de ceux qui ne la changent pas [Klein *et al.*, 2002] : en effet, si une modification d'un libellé préféré de concept terminologique a pour conséquence de changer son sens, l'impact en sera potentiellement très important. En particulier, le changement de conceptualisation peut invalider des relations sémantiques, dont celles définissant l'interopérabilité entre SOC (alignements exacts). Les conséquences sont donc réelles et difficiles à prévoir. À l'inverse, les changements de type « explicatif » ne changent pas la conceptualisation (modification d'une date par exemple) et ne constituent pas de problèmes potentiels quant aux relations sémantiques impliquant le concept.

Par ailleurs, d'autres travaux tendent également à mesurer les impacts structurels et fonctionnels de modifications de concepts : Lee a définit quatre grands types de modifications lors de mises à jour de libellés préférés de la SNOMED CT et a également démontré les éventuels impacts importants lors de modifications de hiérarchies [Lee *et al.*, 2011].

Notre approche initiale, pour le S3M, a d'abord été celle de la mise à jour complète en « écrasant » la version précédente. Les concepts supprimés dans la nouvelle version sont complètement supprimés de la base de données et ainsi, toutes les relations y faisant référence. La philosophie du S3M était alors de ne proposer aux utilisateurs que la version actuelle, et donc valable, de chaque SOC. Cette approche ne peut être satisfaisante à long terme pour tous les types d'utilisateurs pour les raisons expliquées avant (rétro-compatibilité, gestion du SOC, etc.).

L'approche actuelle consiste toujours à conserver une seule version (donc la plus récente) de chaque SOC mais pour toute nouvelle mise à jour, un ensemble de règles permet de conserver un certain nombre d'éléments considérés comme importants alors que d'autres sont purement écrasés (cf. 4.5.3).

Enfin, pour certains cas particuliers comme celui de la CIM, plusieurs versions coexistent dans le S3M en tant que SOC différents. Cependant, aucun lien n'est établit entre les concepts identiques dans les versions disponibles à part un éventuel alignement automatique exact. Par exemple, les CIM-9 et CIM-10 sont intégrées dans le S3M. L'UMLS ou encore BioPortal ont également choisi cette approche pour conserver les différentes versions. Cependant, dans ces systèmes, les versions d'un même SOC sont regroupées. Il est également à noter qu'une grosse différence de gestion de versionnage entre HeTOP et les autres systèmes proches concerne le multi-linguisme. En effet, dans l'UMLS ou dans BioPortal, les variantes par langue d'un SOC sont considérées comme autant de versions de ce SOC. Cela pose évidemment des problèmes de gestion et d'affichage dans les applications.

Enfin, une nouvelle approche de versionnage envisagée, plus technique, consiste à

utiliser l'historisation lors d'édition des SOC par les experts de l'équipe. Cela permettrait une gestion plus fine des modifications.

Versionnage des alignements

La problématique de maintenance des alignements exacts entre SOC a été récemment étudiée par Reis [Reis *et al.*, 2012]. Cette étude affirme d'abord qu'il s'agit d'un sujet encore peu étudié mais crucial. Elle définit ensuite un certain nombre de types de changements de concepts lors de la mise à jour de SOC (fusion, éclatement). Nous pensons nous inspirer de ces travaux et de [Reis *et al.*, 2013] pour mener une étude au sein du S3M afin de mesurer l'impact sur le système.

4.5.3 Mises à jour des SOC dans le S3M

Si une version antérieure d'un SOC est déjà dans le S3M, les concepts déjà présents sont mis à jour (insertion des nouvelles propriétés, écrasement des propriétés existantes). Si des propriétés de concepts sont supprimées dans la nouvelle version, celles-ci sont conservées avec une annotation particulière pour bien les identifier. La même règle s'applique en cas de suppression de concept dans la nouvelle version. Pour les nouveaux concepts, ils sont simplement insérés.
En ce qui concerne les relations hiérarchiques, elles sont complètement écrasées pour éviter d'éventuelles incohérences (cycles, etc.) inter-versions.

4.5.4 Suivi des modifications : historisation du S3M

La nouvelle approche conçue dans le cadre de ces travaux consiste à implémenter l'historisation dans la base de données. Concrètement, il s'agit de dupliquer le modèle physique présenté dans 3.6 et y ajoutant quelques colonnes supplémentaires. On obtient alors un clone dans lequel on va tracer toute modification du schéma original. Une colonne spécifie s'il s'agit d'une insertion, d'une mise à jour ou d'une suppression. On stocke l'ancienne et la nouvelle valeur, le cas échéant. Comme il n'est pas permis de modifier un identifiant (`RDF_RESOURCE`) de ressource du S3M, on peut alors récupérer facilement toute modification pour un identifiant donné et ainsi retracer l'historique de ses éditions. Ce système permet de tracer à la fois les petites modifications ponctuelles et les changements liés au passage à une nouvelle version.

4.5.5 Implémentation

L'implémentation des méthodes d'historisation explicitées précédemment se fera dans les prochains mois. Différents outils de gestion de versions collaboratifs d'ontologies permettent déjà d'élaborer des interfaces graphiques pertinentes pour les

utilisateurs. Parmi ces outils, on peut citer : SemVersion [Voelkel *et al.*, 2006], OntoView [Klein *et al.*, 2002], PROMPTDIFF [Noy *et al.*, 2004]. Concrètement, il s'agit de proposer aux experts du S3M un menu dans l'interface d'édition où seront listés tous les éléments de l'histoire d'un concept, d'un attribut ou d'une relation (date, valeurs, auteur, commentaires, etc.). Il sera alors possible d'éventuellement revenir à une valeur d'une ancienne version ou de comprendre l'origine de certaines modifications.

4.6 Enrichissements des SOC

Tous les SOC intégrés au S3M subissent des traitements d'enrichissements. Tout d'abord, il s'agit de processus automatiques, ensuite, il s'agit de valider (ou invalider) ces informations ajoutées automatiquement (par des experts, via des outils dédiés, cf. 5.5.3). Enfin, selon les SOC, il y a parfois une volonté d'ajouter du contenu manuellement. Il existe plusieurs types d'enrichissement de SOC, qu'ils soient automatiques, semi-automatiques ou manuels :

- Ajout de traductions de PT de concepts
- Ajout de synonymes
- Ajout d'alignements vers d'autres SOC (cf. 2.7)
- Ajout d'attributs spécifiques
- Ajout de relations spécifiques

Cet aspect d'expertise en fonction des SOC intégrés présente une valeur ajoutée très appréciée des utilisateurs mais aussi des éditeurs de SOC. Cependant, comme beaucoup de ces traitements dépendent du réseau sémantique du S3M, des petites erreurs peuvent engendrer des cascades d'erreurs beaucoup plus difficiles à repérer (cf. 6.2.2).

Seuls deux traitements sont exécutés à chaque nouvelle intégration (ou mise à jour) de SOC : la traduction automatique en français ou en anglais des PT et l'alignement exact du SOC intégré vers tous les autres SOC du S3M. Les algorithmes utilisés sont essentiellement basés sur la normalisation des termes (cf. 4.7.1).

4.6.1 Alignements exacts

La génération automatique des alignements exacts repose sur les travaux de Tayeb Merabti [Merabti, 2010]. À chaque nouvelle intégration ou mise à jour de SOC dans le S3M, l'ensemble des nouveaux termes est normalisé puis comparé avec ceux des autres SOC du S3M. Les résultats d'adéquation sont confrontés à une liste de faux-positifs constituée au fil du temps par les experts de l'équipe qui supervisent ces alignements. En effet, via une interface dédiée (cf. 5.5.3), il est possible de valider (ou d'invalider) les alignements générés automatiquement mais aussi de préciser si l'alignement est exact ou approché (plus fin, plus large, ...).

Via l'outil d'édition des SOC, les experts peuvent également ajouter des alignements *de novo*.
Des campagnes de validation d'alignements automatiques et d'ajouts d'alignements exacts sont parfois organisées au sein de l'équipe par les experts. Cela est souvent le cas pour des projets de recherche se focalisant sur tel ou tel SOC (cf. 6.4.2).

4.6.2 Apports de l'UMLS

Le méta-thésaurus UMLS apporte un contenu très important basé sur les CUI. Non seulement ces CUI permettent de retrouver les concepts identiques dans SOC différents (alignements conceptuels) mais ils offrent aussi la possibilité d'utiliser le réseau sémantique de l'UMLS, via les Groupes et les Types Sémantiques (TS). Après l'intégration d'un nouveau SOC dans le S3M qui ne figure pas dans l'UMLS et après la validation d'alignements exacts entre des nouveaux concepts et des concepts existants (et également dans l'UMLS), il est alors possible de leur rattacher des CUI et donc, des TS. Il s'agit ni plus ni moins d'une extension de l'UMLS. Cependant, aucun SOC non inclus dans l'UMLS n'a été entièrement aligné pour adjoindre à chaque concept un CUI ; cela prendrait énormément de temps et de plus, il faudrait créer des nouveaux CUI pour tout concept non retrouvé dans l'UMLS.
Pour éviter d'éventuelles cascades d'erreurs (6.2.2), seuls certains SOC particuliers ont subit ces ajouts pour des contextes de recherche bien particuliers (ATC[5] ou BNPC[6], par exemple).

4.6.3 Relations riches

D'autres relations sont parfois ajoutées entre SOC. On parle parfois de « relations riches » ou « conceptuelles » pour désigner des relations entre concepts possédant une signification précise mais non sémantique (cf. 2.6). Deux exemples sont présentés ci-après pour illustrer des méthodes et des utilisations différentes de ces relations riches ajoutées spécifiquement. Il s'agit encore une fois de contenu à grande valeur-ajoutée et ce genre de traitement serait plus difficile à mettre en œuvre sans un serveur multi-terminologique multi-lingue comme le S3M.

5. Anatomical Therapeutic Chemical classification, éditée par l'OMS : http://www.whocc.no/atc_ddd_index/

6. Base Nationale des Produits et des Compositions, éditée par le Centre antipoison et de toxicovigilance du CHU de Nancy : http://www.sante-environnement-travail.fr/minisite.php3?id_rubrique=1011&id_article=4346

Expansion des phénotypes OMIM de HPO vers d'autres SOC

La *Human Phenotype Ontology* (HPO)[7] a pour but de fournir un vocabulaire standard des anormalités phénotypiques engendrées par les maladies chez l'Homme. Originellement développée à partir de la base Online Mendelian Inheritance in Man (OMIM)[8], HPO lie des phénotypes aux maladies présentes dans OMIM mais également aux gènes de la base *Gene* du NCBI[9].
Afin d'améliorer HPO et les SOC où l'on peut trouver des maladies, le réseau sémantique a été utilisé pour rattacher directement les maladies hors HPO alignées exactement à des maladies de OMIM aux termes HPO correspondants. Par exemple, le phénotype HPO « Dolichosténomélie » (HPO :0001519) est lié nativement (entre autres), avec le « Syndrome de Beals » de OMIM (121050). Or, ce code OMIM est aligné exactement avec la maladie HRDO[10] « Arachnodactylie congénitale avec contractures » (115) puisqu'il s'agit, en réalité, de la même maladie. On peut alors relier ce code HRDO avec le phénotype HPO par une relation de type « A pour Phénotype HPO ». On obtient alors une liste de phénotypes pour les maladies concernées (Figure 4.3) ainsi qu'une liste étendue de maladies non OMIM pour chaque concept HPO (Figure 4.4).
Ces traitements, ainsi que d'autres travaux sur HPO, ont d'ailleurs fait l'objet d'une publication [Grosjean *et al.*, 2013]. Le nombre total de relations entre phénotypes HPO et maladies (hors OMIM) ajoutées est de 71 586 ce qui correspond à 6 559 phénotypes HPO distincts (soit 55% des phénotypes HPO).

7. Éditée par le groupe CBB de l'Institut pour la Génétique Médicale et Humaine de l'université de médecine Charité à Berlin, Allemagne : http://www.human-phenotype-ontology.org/contao/index.php/hpo_home.html
8. Éditée par l'Institut McKusick-Nathans de Médecine Génétique, Baltimore, Maryland, USA : http://www.ncbi.nlm.nih.gov/omim
9. National Center for Biotechnology Information, Washington DC, Maryland, USA : http://www.ncbi.nlm.nih.gov/gene/
10. Human Rare Diseases Ontology, dérivée de Orphanet, classification des maladies rares : www.orpha.net/

FIGURE 4.3 – Capture d'écran de HeTOP des relations de la maladie HRDO « Arachnodactylie congénitale avec contractures » : liste des phénotypes HPO déduite grâce au réseau sémantique

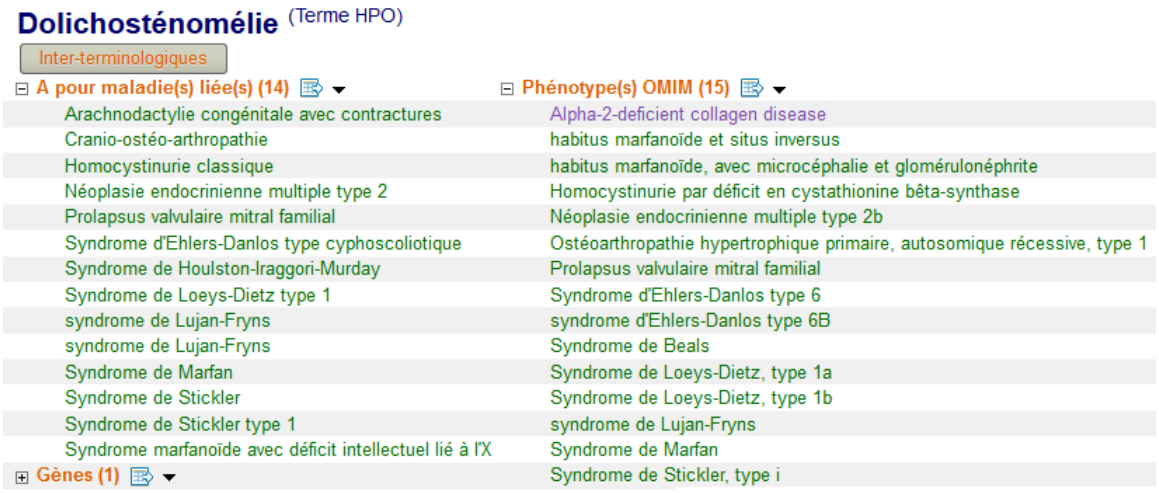

FIGURE 4.4 – Capture d'écran de HeTOP des relations du concept HPO « Dolichosténomélie » : listes des maladies OMIM et des autres maladies du S3M déduite grâce au réseau sémantique

Co-occurrences de dispositifs médicaux LPP et d'actes médicaux CCAM

La LPP est la Liste des Produits et Prestations remboursables par l'Assurance Maladie. Il s'agit notamment des dispositifs médicaux pour traitements et matériels d'aide à la vie, aliments diététiques et articles pour pansements, des orthèses et prothèses externes, des dispositifs médicaux implantables et des véhicules pour handicapés physiques [11].

La Classification Commune des Actes Médicaux (CCAM) est une nomenclature française destinée à coder les gestes pratiqués par les médecins, gestes techniques dans un premier temps puis, par la suite, les actes intellectuels cliniques. Cette classification sert notamment, dans les hôpitaux publics et privés, à coder les dossiers des patients dans le Programme de Médicalisation des Systèmes d'Information (PMSI). Ces deux classifications sont souvent utilisées conjointement dans des Systèmes d'Information Hospitaliers (SIH) en France pour assurer une tarification des actes et une traçabilité sanitaire. De fait, il est possible, en détectant des co-occurrences, de lier certains dispositifs LPP à des actes médicaux CCAM. Un travail spécifique, inspiré de [Carpentier *et al.*, 2010], a été fait dans ce sens au CHU de Rouen pour constituer des relations entre dispositifs LPP et actes CCAM reparties en deux types : les dispositifs requis pour un acte et ceux utiles pour un acte. L'étude a porté sur des données de patients du CHU de Rouen entre 2004 et 2009 soit environ 47 000 actes. Une matrice de co-occurrences entre les actes et les dispositifs a été filtrée automatiquement en fonction des fréquences et enfin manuellement. Cela a aboutit à la création de 5 357 relations de type « requis pour » et 1 272 relations de type « utile pour ».

Ce travail peut donner lieu à deux utilisations possibles dans les SIH : l'aide à la saisie lors du codage afin de faciliter la sélection des codes mais également au contrôle lors de cette saisie pour éviter des erreurs humaines lorsque le code LPP entré n'est pas en adéquation avec l'acte ou vice et versa.

Ces relations ont été intégrées dans le S3M et sont visibles notamment dans HeTOP : on peut voir sur la Figure 4.5 la liste des dispositifs LPP utiles ou requis pour l'acte de « Remplacement du disque intervertébral par prothèse ». Sur la Figure 4.6), on peut visualiser les relations inverses reliant ces actes CCAM à un dispositif LPP, ici « Rachis, coussinet ».

11. http://www.codage.ext.cnamts.fr/codif/tips/index_presentation.php?p_site=AMELI

FIGURE 4.5 – Capture d'écran de HeTOP des relations de la fiche LPP « Rachis, coussinet » : liste des actes CCAM qui requièrent ce dispositif

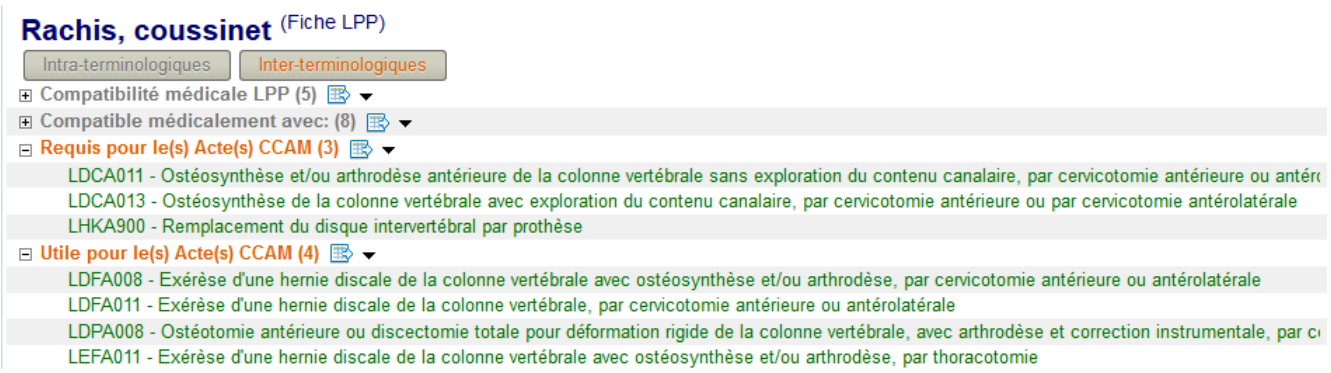

FIGURE 4.6 – Capture d'écran de HeTOP des relations de l'acte CCAM « Remplacement du disque intervertébral par prothèse » : listes des dispositifs LPP utiles ou requis par cet acte

4.6.4 Attributs spécifiques

D'autres propriétés de concepts terminologiques peuvent être ajoutées aux SOC intégrés dans le S3M. Les attributs sont des données souvent textuelles ou numériques correspondant à autant de méta-données disponibles pour les SOC. Les synonymes, CUI UMLS, annotations ou les définitions sont également considérés comme des attributs. L'exemple suivant présente l'intégration dans le S3M d'attributs particuliers visant à représenter graphiquement les concepts terminologiques grâce à une ou plusieurs icônes.

Intégration du langage iconique VCM dans le S3M

Le langage iconique VCM (Visualisation de Connaissances Médicales) a pour objectifs de représenter des signes, des maladies, des états physiologiques, des habitudes de vie, des médicaments ou des examens para-cliniques afin d'accéder rapidement à une information [Lamy *et al.*, 2008]. Développé dans le cadre de la thèse de Jean-Baptiste Lamy[Lamy, 2006] puis du projet de recherche ANR L3IM [12], ce langage a d'abord été utilisé pour représenter des concepts terminologiques dans les SOC suivants : MeSH, CIM-10, ATC (médicaments) puis une partie de la SNOMED CT [13] (axe des maladies). En terme d'utilisation, ces icônes peuvent servir notamment à visualiser rapidement le contenu de ressources indexées voire même à filtrer les résultats de recherches [Lamy *et al.*, 2010].
L'équipe CISMeF a participé à L3IM en réalisant, entre autres, la tâche d'alignement entre le langage VCM et les autres SOC du projet. VCM est lui-même représenté en tant qu'ontologie [Lamy *et al.*, 2013] et a donc été intégré dans le S3M. Il est composé de « primitives » réparties en sept composantes. Les combinaisons de ces primitives constituent les icônes. Par exemple, la combinaison « en_cours–pathotumeurb–vesicule–rien–rien–rien–rien » désigne une tumeur bénigne au niveau de la vésicule ou des voies biliaires (Figure 4.7). Les alignements effectués manuellement ont été hérités pour tous les concepts terminologiques situés aux niveaux hiérarchiques inférieures (« explosion hiérarchique »). Cela évite de devoir tout aligner et donc, fait gagner beaucoup de temps. L'explosion hiérarchique des icônes VCM peut générer des redondances ou des conflits d'icônes. Cela est évité grâce à des algorithmes (développés avec le langage) qui permettent de fusionner ou de filtrer des icônes. Enfin, les icônes VCM ont été « poussés » vers d'autres SOC via le réseau sémantique d'alignements exacts.
Grâce aux traitements manuels puis automatiques, 228 030 icônes VCM décrivent 179 893 concepts terminologiques différents dans le S3M (un concept peut être représenté par plusieurs icônes). 93% des codes ATC, 41% des descripteurs MeSH et

12. *Langage Iconique et Interfaces Interactives en Médecine* - ANR TecSan 2008 : http://projet4-limbio.smbh.univ-paris13.fr/
13. Systematized Nomenclature of MEDicine Clinical Terms, éditée par la *Health Terminology Standards Development Organisation (IHTSDO)*

38,5% de la CIM-10 possèdent au moins une icône. La Figure 4.8 illustre l'affichage des icônes VCM pour un concept donné dans HeTOP. Une des études menées en

FIGURE 4.7 – Capture d'écran de HeTOP des relations vers les primitives VCM pour le descripteur MeSH « Kyste du cholédoque »

FIGURE 4.8 – Capture d'écran de HeTOP d'un exemple d'icône VCM pour le concept SNOMED CT « Diverticule congénital de l'estomac »

marge de ce projet a été d'évaluer la qualité des alignements entre ce langage iconique et les terminologies choisies. Les très bons résultats publiés dans [Griffon *et al.*, 2014b] ont permis de valider l'approche et d'intégrer les icônes créés dans HeTOP. Une autre étude a été menée par la suite pour évaluer l'apport de ce langage dans l'utilisation d'outils intégrant les icônes : celle-ci révèle un apport significatif de ce système dans le moteur de recherche Doc'CISMeF [Griffon *et al.*, 2014a].

4.7 Pré- ou post-traitements des SOC

Afin d'être utilisées rapidement et efficacement par les applications, les données de SOC nécessitent certains traitements automatiques, une fois intégrées dans le S3M. Ces traitements sont appelés communément « batchs » (terme issu de l'anglais, « traitement par lots » en français) et sont des appels à différents programmes effectuant des tâches automatiques, et ce, de façon régulière. Ainsi, différentes étapes sont nécessaires en fonction des types de données (attributs, relations hiérarchiques, alignements, etc.).

4.7.1 Normalisations

Pour pouvoir être exploités dans des contextes de RI ou d'alignements, les libellés doivent subir des traitements de normalisation. Au sens du Traitement Automatique de la Langue (TAL), ces normalisations consistent en une série d'opérations afin d'obtenir des libellés directement comparables. Voici le détail des étapes de normalisation effectuées dans le S3M (exemple en Figure 4.9) :

1. Dés accentuation
2. Suppression de caractères spéciaux
3. Suppression d'espaces inutiles
4. Uniformisation de la casse (tout en minuscule)
5. Réduction de chaque mot à sa racine (racinisation ou « stemming » en anglais)
6. Mots normalisés mis dans l'ordre alphabétique

```
Terme original : Ischémie de la moelle épinière
1/2/3. : Ischemie de la moelle epiniere
4. : ischemie de la moelle epiniere
5. : ischem;moel;epinier;
6. : epinier;ischem;moel;
```

FIGURE 4.9 – Exemple de normalisation dans le S3M pour un terme

Ces libellés normalisés sont stockés dans la base de données du S3M pour chaque terme considéré comme important dans un contexte de RI. À chaque insertion ou mise à jour d'un libellé dans le S3M, un programme met à jour les valeurs normalisées. Trois valeurs sont en fait calculées correspondant aux étapes 4, 6 et 6 sans racinisation (utilité différente selon les programmes).

4.7.2 Hiérarchies

La structure hiérarchique terminologique est représentée sous forme d'arbre orienté. Certains SOC sont dits « polyhiérarchiques », c'est-à-dire qu'un concept peut être

représenté dans le même arbre mais avec des « chemins » différents. Un chemin hiérarchique est donc défini par l'ensemble des identifiants de concepts permettant de lier le plus haut niveau de l'arborescence à un concept donné. L'arbre hiérarchique est stocké dans la base de données avec un couple « concept père » - « concept fils » par ligne. Dans l'univers des SOC, ces concepts sont appelés respectivement « Broader Term » (BT, plus large) et « Narrower Term » (NT, plus étroit). Lorsque l'arbre est « polyhiérarchique », on adjoint à ce couple BT-NT, un troisième élément pour spécifier son chemin original unique.
Pour que les applications puissent directement afficher et exploiter ces arbres, il est nécessaire de calculer les chemins formellement. Pour cela, lorsque le SOC ne propose pas nativement ces chemins dans les données sources, un algorithme dédié a été mis en place. Le principe est de parcourir l'arbre via une requête SQL récursive et d'associer à chaque couple BT-NT un chemin construit pas-à-pas par concaténations. Par ailleurs, il est également nécessaire d'ajouter à chaque chemin un élément pour savoir si un concept terminologique est une feuille dans l'arbre ; une feuille étant un élément de l'arbre n'ayant pas de fils, il faut pouvoir afficher qu'il s'agit donc du niveau le plus bas de l'arbre. Dans ce cas, le chemin calculé sera suffixé par un caractère dédié (« ! »). Ainsi, par exemple, la représentation en triplets hiérarchiques du couple BT-NT des descripteurs MeSH « Kyste du cholédoque » (D015529) et « Maladie de Caroli » (D016767) est affichée dans le Tableau 4.1. Sa représentation graphique sous forme « d'arbre-tableau » est illustrée dans HeTOP sur la Figure 4.10. On voit bien sur cet exemple qu'un seul couple BT-NT peut être défini par plusieurs chemins. Enfin, pour représenter le plus proprement possible ces arbres et avoir des chemins cohérents, il est souvent nécessaire d'ajouter un concept terminologique correspondant au plus haut niveau de la hiérarchie.
Lors de la mise à jour d'un SOC dans le S3M, l'ensemble de la hiérarchie du SOC

BT	NT	Chemin
D015529	D016767	ARBO/C/D009358/D000013/D004065/D015529/D016767!
D015529	D016767	ARBO/C/D004066/D001660/D001649/D015529/D016767!
D015529	D016767	ARBO/C/D004066/D004065/D015529/D016767!

TABLE 4.1 – Représentation simplifiée des triplets hiérarchiques dans le S3M (table `TB_HIERARCHY`)

est écrasé et une procédure recalcule tous les chemins possibles. Ce traitement peut prendre de plusieurs secondes à plusieurs minutes, en fonction du nombre de concepts dans la hiérarchie.

Problème de l'édition des hiérarchies

Comme elle n'est effectuée qu'à chaque nouvelle insertion ou mise à jour de SOC, cette procédure n'a pas d'impact pour les utilisateurs des applications du

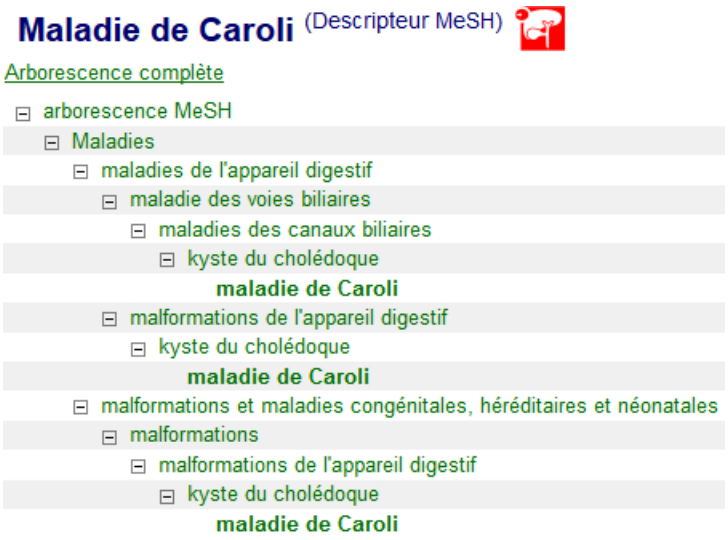

FIGURE 4.10 – Capture d'écran de HeTOP d'un exemple de représentation hiérarchique du descripteur MeSH « Maladie de Caroli »

S3M. Cependant, si l'on voulait éditer les hiérarchies en changeant un concept de place, en ajoutant un nouveau concept à une hiérarchie ou en supprimant un concept, cela poserait un problème évident : en fonction du ou des endroits de la hiérarchie impactée par la modification, recalculer tous les chemins en temps réel serait trop long pour un utilisateur. En effet, si on intercale un concept à très haut niveau d'une hiérarchie par exemple, tous les chemins des couples situés en dessous serait impactés et donc, à recalculer. Il s'agit d'un problème technique encore irrésolu dans le S3M. Pour l'instant, cela n'a pas d'importance puisqu'il n'est encore jamais arrivé qu'il faille modifier une hiérarchie. Cependant, pour la création de nouvelles terminologies par exemple (cf. 5.5.4), cela pourrait poser problème. En attendant de développer une méthode plus efficace, les changements de hiérarchies ne peuvent être faits en temps réel.

4.7.3 Réseau sémantique

Il existe quatre types de processus automatiques impliquant le réseau sémantique :

- Le premier est celui qui crée de nouveaux alignements exacts entre les concepts d'un nouveau SOC intégré et ceux des SOC existants (4.6.1).
- Le deuxième est celui d'ajout de nouveaux alignements conceptuels basés sur l'UMLS (4.6.2).
- Le troisième concerne d'éventuels attributs à « pousser » via les alignements exacts validés, comme par exemple les icônes VCM (cf. 4.6.4).
- Enfin, le dernier traitement consiste valider automatiquement certains alignements par transitivité. En mathématiques, une relation transitive est une relation binaire pour laquelle une suite d'objets reliés consécutivement aboutit à

une relation entre le premier et le dernier. Ainsi, les alignements exacts (donc d'équivalence) entre concepts terminologiques sont transitifs (c'est également le cas pour les relations « is-a » ou « part-of » des hiérarchies mais ce n'est pas le cas pour les alignements considérés comme « faux » par exemple). En effet, si α est aligné exactement avec β et que β est aligné exactement avec γ alors α peut être aligné exactement avec γ. La Figure 4.11 illustre cette transitivité avec trois concepts terminologiques. Ce processus de validation automatique est très pratique puisqu'il permet de faire gagner du temps aux experts responsables des validations des alignements exacts automatiques et des nouveaux alignements exacts manuels.

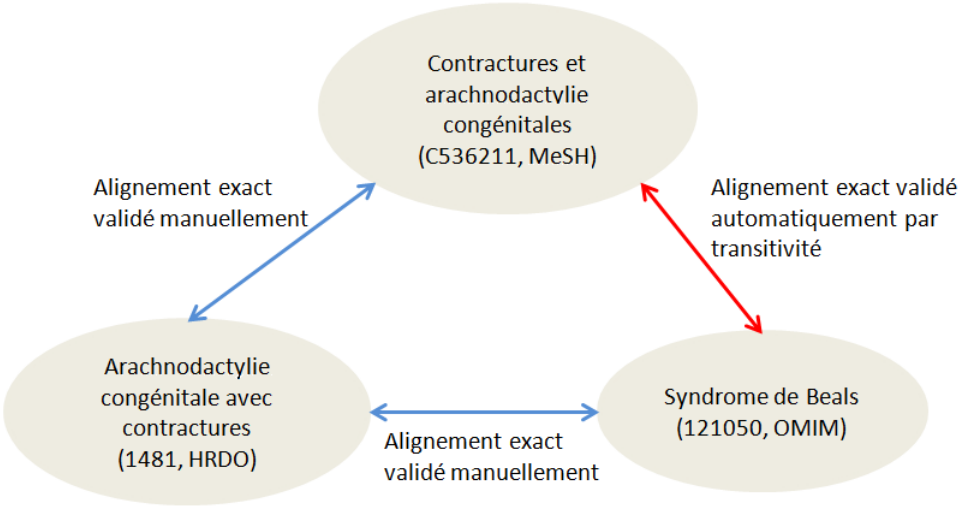

FIGURE 4.11 – Exemple de validation automatique d'alignement exact par transitivité dans le S3M

4.8 Exemples d'intégration

Afin d'illustrer au mieux l'application des méthodologies développées, quelques exemples d'intégration de SOC sont détaillés ci-après. Un cas concerne l'intégration d'un SOC « promu » (NABM), un autre celle d'une vraie terminologie (SNOMED 3.5 internationale) et le dernier cas concerne l'intégration d'une ontologie (FMA).

4.8.1 Intégration d'un SOC « promu » : la NABM

Dans le cas d'un SOC « promu », il faut créer un modèle terminologique à partir d'une éventuelle documentation ou d'une expertise quelconque. En dehors du fait de trouver une source de données correcte exploitable automatiquement, l'un des soucis majeurs de la promotion d'un SOC vers un cadre terminologique est celui de la compréhension du SOC lui-même. La façon dont le fichier source est structuré, sa documentation, son utilisation et l'avis d'experts sont autant d'éléments clés dans

le choix d'une modélisation pour un SOC (cf. 2.5).
Dans le cas de la NABM, un certain nombre de documents officiels sont disponibles ainsi qu'un site web dédié et son utilisation est définie clairement.
La NABM (Nomenclature des Actes de Biologie Médicale) établit la liste des actes susceptibles d'être pris en charge par l'assurance maladie et leur cotation exprimée en lettre-clé B. Cette nomenclature s'impose aux prescripteurs en ce qui concerne le respect des indications qui conditionnent la prise en charge et aux directeurs de laboratoire, notamment en ce qui concerne le respect des obligations techniques particulières et la facturation des actes. La NABM est disponible en ligne sur le site AMELI (Assurance Maladie en Ligne)[14], dans la partie dédiée aux nomenclatures nationales gérées par cette structure (avec la CCAM et la LPP)[15]. Un guide officiel détaillé est également disponible[16] et constitue un cadre extrêmement précis sur la façon dont la nomenclature a été créée et comment il est nécessaire de l'utiliser. La création d'un modèle terminologique d'une telle nomenclature permet de structurer le vocabulaire mais ne se substitue absolument pas au guide d'utilisation. En d'autres termes, l'intégration de ce type de SOC dans le S3M ne permet pas aux utilisateurs potentiels de comprendre directement le fonctionnement du SOC et encore moins des bonnes pratiques d'utilisation.

CODE	CHAPITRE	SOUS-CHAPITRE	COEFFICIENT B	DATE CREATION	LIBELLE	ENTENTE PREALABLE	REMBOURSEMENT 100%	NBR MAXI DE CODE	N° REGLE SPECIFIQUE	REF INDICATION MEDICALE	ACTES RESERVES	INITIATIVE BIOLOGISTE	CONTINGENCE TECHNIQUE	R.M.O.	EXAMEN SANGUIN	DERNIERE DATE EFFET	CODES INCOMPATIBLES
4084	0	0	500	22/10/2010	DETERMINATION PRENATALE DU SEXE FOETAL SANG MATERNEL		1	1		1	1		3		1	15/03/2011	
9905	0	0	5	14/02/1997	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 5 (SANG)							1			1	19/01/2010	
9910	0	0	10	14/02/1997	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 10 (SANG)							1			1	19/01/2010	
9915	0	0	15	14/02/1997	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 15 (SANG)							1			1	19/01/2010	
9916	0	0	1	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 1 (SANG)							1			1	19/01/2010	
9917	0	0	2	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 2 (SANG)							1			1	19/01/2010	
9918	0	0	3	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 3 (SANG)							1			1	19/01/2010	
9919	0	0	4	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 4 (SANG)							1			1	19/01/2010	
9920	0	0	6	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 6 (SANG)							1			1	19/01/2010	
9921	0	0	7	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 7 (SANG)							1			1	19/01/2010	
9922	0	0	8	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 8 (SANG)							1			1	19/01/2010	
9923	0	0	9	23/12/2009	COMPLEMENT A LA COTATION MINIMALE DE VALEUR B 9 (SANG)							1			1	19/01/2010	

FIGURE 4.12 – Extrait du fichier source de la NABM (format Excel)

Modélisation

En analysant donc le guide officiel de la NABM et son fichier source au format Excel (voir Figure 4.12), on distingue 3 niveaux hiérarchiques : codes, sous-chapitres et chapitres. Plusieurs d'attributs sont disponibles pour chaque code : « coefficient B », « date de création », etc. Il existe aussi une méta-donnée reliant les codes entre

14. http://www.ameli.fr/
15. http://www.codage.ext.cnamts.fr/codif/nabm/index_presentation.php?p_site=AMELI
16. http://www.codage.ext.cnamts.fr/f_mediam/fo/nabm/DOC.pdf

eux. On définit alors la relation du type « Codes incompatibles ». Quelque soit le SOC à intégrer, une expertise est faite sur la pertinence des méta-données à conserver dans le S3M. En effet, le S3M n'a pas pour vocation de se supplanter complètement les portails originaux des SOC. Certains attributs (typiquement, la facturation dans NABM) évoluent rapidement et constituent des informations critiques. Il est souvent nécessaire de filtrer ces attributs pour en extraire les informations pertinentes dans le cadre d'un serveur multi-terminologique essentiellement axé sur l'indexation/codage, la traduction, l'interopérabilité et la RI. La lisibilité et la compréhension par les utilisateurs n'en sont que meilleures.
Ainsi, dans le cas de la NABM, les experts (médecins et biologistes attachés à l'équipe CISMeF, spécialistes des nomenclatures) ont choisi de ne conserver que les codes et leurs relations d'incompatibilité. Le modèle conçu pour la NABM dans le S3M est présenté dans le Figure 4.13. Pour créer le modèle formellement, la première

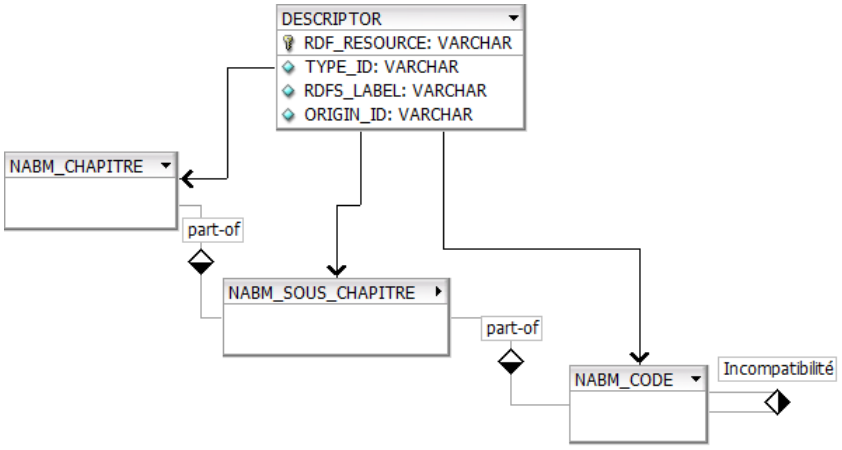

FIGURE 4.13 – Modèle terminologique de la NABM au sein du S3M

représentation se fait sous forme d'objets Java dans le parseur dédié à l'intégration de cette source de données.

Intégration

Un parseur Java est donc écrit spécifiquement pour le format du fichier source de la NABM : il s'agit d'un parseur Java Excel existant (classe fille) dans P1 (décrit plus haut). L'API Java JXL [17] permet d'ouvrir et de parcourir facilement des fichiers Excel. Parallèlement à cette classe Java, une autre classe est écrite où sont répertoriés les différents types de concepts, méta-données (attributs et relations) définissant le modèle du SOC. La classe Java parcourant le fichier Excel va donc utiliser la classe du modèle pour instancier des objets Java représentant les concepts et leurs relations en RDF/XML dans le fichier de sortie. La Figure 4.14 présente un extrait du fichier de sortie du P1 pour la NABM. L'étape suivante consiste à vérifier

17. http://jexcelapi.sourceforge.net/

```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:itm="http://www.mondeca.com/system/itm#"
    xmlns:skos="http://www.w3.org/2004/02/skos/core#"
    xmlns:publishing="http://www.mondeca.com/system/publishing#"
    xmlns:skosm="http://www.w3.org/2004/02/skos/mapping#"
    xmlns:t3="http://www.mondeca.com/system/t3#"
    xmlns:smts="http://www.chu-rouen.fr/smts#"
    xmlns:basicontology="http://www.mondeca.com/system/basicontology#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
  <publishing:BT-NT rdf:about="http://www.chu-rouen.fr/smts#NABM_BTNT~NABM_SC76-NABM_3747">
    <publishing:NT>
      <smts:NABMcode rdf:about="http://www.chu-rouen.fr/smts#NABM_3747">
        <skos:notation>3747</skos:notation>
        <rdfs:label xml:lang="fr">HERPES SIMPLEX (VHS) : SD : AC IGM ANTI VHS 1 + 2 PAR IFI + ITERATIF</rdfs:label>
      </smts:NABMcode>
    </publishing:NT>
    <publishing:BT>
      <smts:NABMsousChapitre rdf:about="http://www.chu-rouen.fr/smts#NABM_SC76">
        <skos:notation>SC76</skos:notation>
        <rdfs:label xml:lang="fr">Sous-chapitre 6</rdfs:label>
      </smts:NABMsousChapitre>
    </publishing:BT>
  </publishing:BT-NT>
  <smts:NABMincompatibiliteRelation rdf:about="http://www.chu-rouen.fr/smts#NABM_NABMincompatibilite~NABM_1425-NABM_1
    <smts:NABMestIncompatible2>
      <smts:NABMcode rdf:about="http://www.chu-rouen.fr/smts#NABM_1426">
        <skos:notation>1426</skos:notation>
        <rdfs:label xml:lang="fr">TOXOPLASMOSE CAS GENERAL : CULTURE ET INOCULATION</rdfs:label>
      </smts:NABMcode>
    </smts:NABMestIncompatible2>
    <smts:NABMestIncompatible1>
      <smts:NABMcode rdf:about="http://www.chu-rouen.fr/smts#NABM_1425">
        <skos:notation>1425</skos:notation>
        <rdfs:label xml:lang="fr">TOXOPLASMOSE CAS GENERAL :  INOCULATION SOURIS</rdfs:label>
      </smts:NABMcode>
    </smts:NABMestIncompatible1>
  </smts:NABMincompatibiliteRelation>
```

FIGURE 4.14 – Extrait du fichier de sortie de P1 en RDF/XML pour la NABM

l'intégrité des données extraites via l'Outil SMTS présenté précédemment. Cette application va permettre de compter les concepts extraits et les relations mais aussi de les séparer dans deux fichiers distincts. Elle va aussi générer un fichier correspondant au modèle terminologique déduit à partir des données. Il s'agit d'un fichier OWL-Lite auquel un certains nombres d'éléments ont été ajoutés automatiquement ou à la main (contraintes, libellés de méta-données, ...). Ces informations sont utiles pour l'intégration dans la base de données.

En effet, la phase suivante est celle de l'intégration proprement dite dans la base de données via le P2. Ce programme s'appuie sur le fichier du modèle pour détecter les types de concepts, les relations et autres méta-données pour insérer les informations dans la base. Là encore, un contrôle est possible car toute erreur potentielle est affichée dans une console. À la fin de l'intégration, un bilan chiffré avec les nombres d'insertions, d'erreurs ou de données éventuellement déjà existantes apparait (cf. Figure 4.15).

Après l'intégration des données, il reste encore quelques petites étapes à effectuer : génération des chemins hiérarchiques, des libellés normalisés, des alignements automatiques vers les autres SOC du S3M, réglages de l'affichage des méta-données et des liens contextuels (pour la NABM, un lien vers le site AMELI) pour les applications (partie MODEL via le DBGUI, cf. 5.5.1). La plupart de ces opérations sont automatiques mais nécessitent un contrôle.

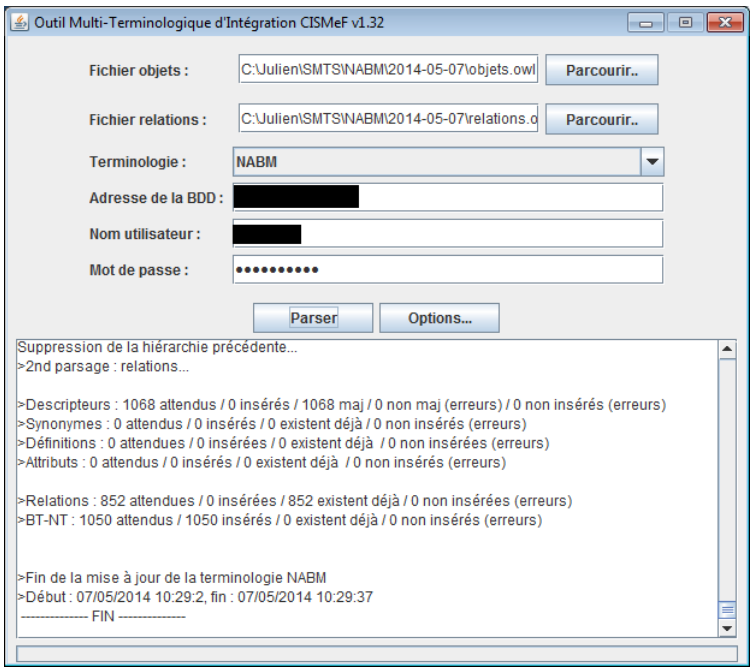

FIGURE 4.15 – Capture d'écran de l'application P2 pour la NABM v41

Enfin, la NABM est prête à être consultée directement dans les différentes applications du SI exploitant le S3M. La Figure 4.16 est une capture d'écran de HeTOP sur la hiérarchie de la NABM développée pour le code « FECONDATION IN VITRO PAR MICROMANIPULATION (ICSI) » (0061).

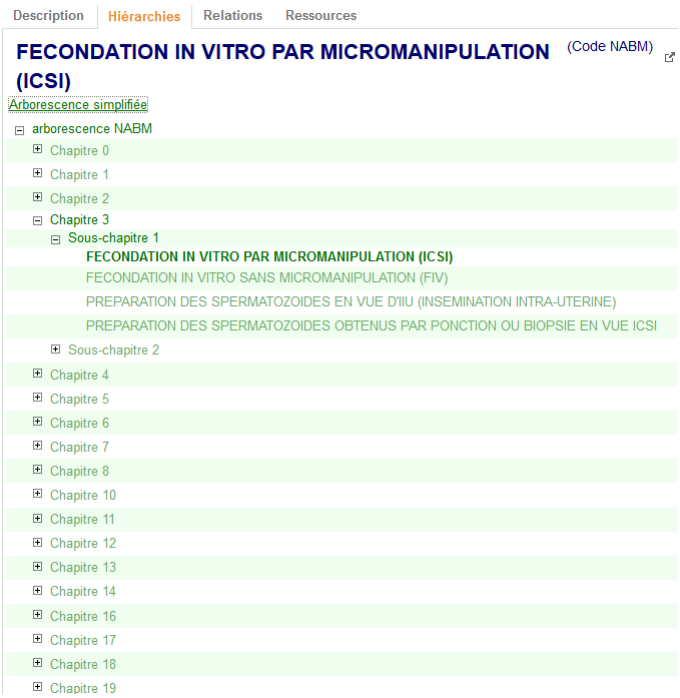

FIGURE 4.16 – Capture d'écran de HeTOP pour la hiérarchie développée du code 0061 de la NABM

4.8.2 Intégration d'une terminologie native : la SNOMED 3.5 (internationale)

L'intégration d'une terminologie native, c'est-à-dire déjà modélisée, est très proche de la démarche décrite précédemment. Cependant, la première phase d'expertise sur la création du modèle est alors inutile. Il est tout de même parfois nécessaire de bien lire la documentation pour récupérer le modèle terminologique qui n'est pas toujours formellement décrit et disponible dans un fichier électronique.
La SNOMED internationale (version 3.5) est une classification pluri-axiale couvrant tous les champs de la médecine et de la dentisterie humaines, ainsi que la médecine animale. Il s'agit d'un système de classification permettant de normaliser l'ensemble des termes médicaux utilisés par les praticiens de santé. La SNOMED a pour fonction d'attribuer un code à chaque concept permettant un grand nombre de combinaisons entre eux. Elle comprend également une liste des diagnostics interfacée avec la CIM-10. La SNOMED permet ainsi de stocker des informations médicales individuelles dans des entrepôts de données afin d'établir des outils d'analyse décisionnelle, de faciliter des décisions thérapeutiques, de contribuer aux études épidémiologiques et à l'enseignement. L'utilisation de ce SOC garantit, théoriquement, l'universalité du vocabulaire médical.
La SNOMED 3.5 est disponible sous plusieurs formats et dans plusieurs langues. Côté français, l'ASIP Santé[18] administre une version officielle de cette classification. La version anglaise est disponible notamment directement dans l'UMLS. La SNOMED 3.5 n'est pas une terminologie proprement dite (en tous cas, elle n'est pas définie de la sorte) mais elle possède tout de même une structure et un modèle typiques d'une terminologie. La Figure 4.17 présente ce modèle au sein du S3M. L'intégration de la SNOMED internationale dans le S3M s'est faite sensiblement de la même façon que celle de la NABM étant donné que le format source en français est également un fichier Excel. Cependant, comme la SNOMED internationale est présente dans l'UMLS, un certain nombre de traitements ont été faits à la suite de l'intégration. Il a alors été possible de récupérer des traductions anglaises des libellés préférés mais également des synonymes. De plus, les CUI ont été intégrés ainsi que les alignements conceptuels vers d'autres SOC du S3M ayant également des CUI. Enfin, les types sémantiques ont été reliés aux concepts. À chacune de ces opérations correspond une requête SQL de type « SELECT » dans la base de données de l'UMLS permettant de générer des ordres SQL « INSERT » dans le S3M.

18. L'Agence des Systèmes d'Information Partagés de Santé est une institution publique française qui a pour vocation de renforcer la maîtrise d'ouvrage publique des systèmes d'information se développant dans le secteur de la santé, notamment via les technologies numériques (http://esante.gouv.fr/)

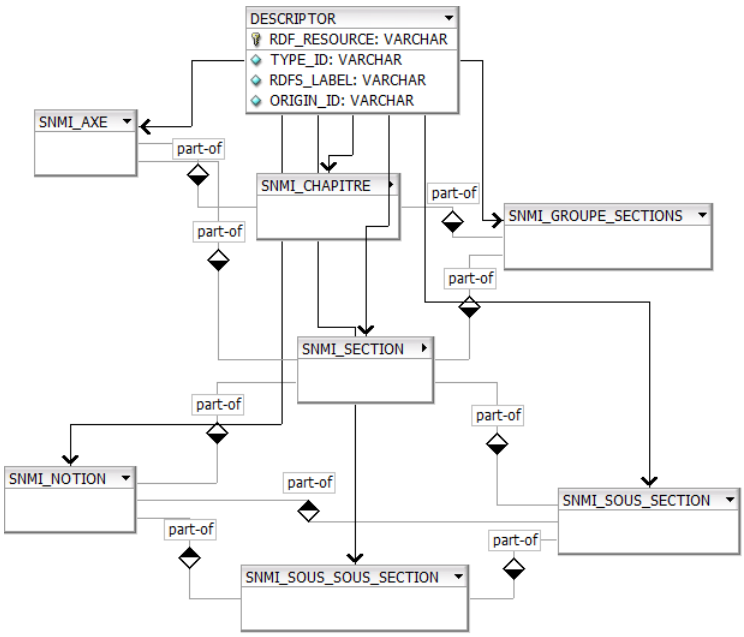

FIGURE 4.17 – Modèle terminologique de la SNOMED 3.5 (internationale) au sein du S3M

4.8.3 Intégration d'une ontologie : la FMA

La *Foundational Model of Anatomy* (FMA)[19] est une source d'évolution des connaissances pour l'informatique biomédicale. Elle représente des catégories ou types et les relations nécessaires à la représentation symbolique de la structure phénotypique du corps humain sous une forme qui est compréhensible pour l'homme et est également navigable, analysable et interprétable par les systèmes informatiques [Rosse & Mejino, 2003]. Plus précisément, la FMA est une ontologie de domaine qui représente un ensemble cohérent de connaissances déclaratives explicites sur l'anatomie humaine. Elle est éditée par le *Structural Informatics Group*, Université de Washington, aux États-Unis.

Dans le cadre d'une collaboration avec le Pr Christine Golbreich en 2010, j'ai d'abord travaillé sur la migration de la FMA en OWL2 [Golbreich *et al.*, 2011], [Golbreich *et al.*, 2013].

Le fichier source de la FMA utilisé pour l'intégration dans le S3M a donc été un fichier OWL issu de ces travaux. Un parseur OWL basé sur la OWL API[20] a été écrit pour l'occasion. Il a ensuite été spécifié pour la FMA (classe Java fille).

Modélisation

La modélisation en une terminologie d'une ontologie formelle est une problématique importante. Il s'agit en effet de « dégrader » un modèle extrêmement complexe, avec des règles et des fonctions difficilement représentables dans un modèle

19. http://sig.biostr.washington.edu/projects/fm/AboutFM.html
20. http://owlapi.sourceforge.net/

terminologique. Ainsi, il est nécessaire de s'abstraire d'un certain nombre d'éléments ontologiques. La méthodologie adoptée dans ces travaux de thèse pour passer d'un modèle ontologique à un modèle terminologique est la suivante : un seul type de concept est créé par ontologie. On obtient alors un modèle très simple mais très peu expressif. Dans le cas de la FMA, le type de concept « Entité FMA » a été créé et tous les concepts terminologiques représentant les classes ontologiques deviennent alors des instances de ces classes (Figure 4.18). En ce qui concerne les propriétés, elles peuvent être conservées telles quelles en tant que relations ou attributs mais la plupart des axiomes de propriétés ne sont pas gardés (symétrie, transitivité, etc.). Cette approche est tout à fait similaire à celle de l'UMLS qui intègre en son sein des ontologies formelles représentées dans une base de données relationnelle : par exemple, la SNOMED CT mais aussi la FMA depuis 2012.

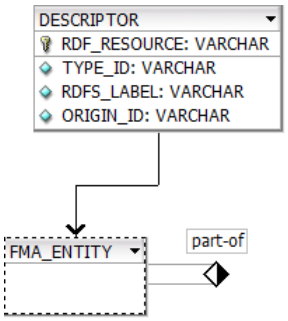

FIGURE 4.18 – Modèle terminologique de la FMA au sein du S3M

Intégration

Une fois le parseur et le modèle écrits, le reste de l'intégration se fait de manière classique. Cependant, comme pour la SNOMED internationale, la FMA étant présente dans l'UMLS, les traitements spécifiques présentés précédemment sont également effectués.
La FMA, contrairement à la NABM ou à la SNOMED internationale, a subi par la suite des ajouts supplémentaires à ceux effectués classiquement (alignements automatiques, etc.) : traductions manuelles et semi-automatiques en français, nouveaux synonymes français, définitions et traductions précises des relations sémantiques à l'aide d'un médecin spécialiste en anatomie (Dr Jean-Michel Müller, CHU de Rouen) [Merabti *et al.*, 2011]. Tout ceci constitue une valeur ajoutée précieuse pour les utilisateurs (étudiants, traducteurs, etc.). Ces ajouts ont d'ailleurs été soumis à l'éditeur de la FMA pour les intégrer dans la prochaine version officielle de l'ontologie.

4.9 Synthèse du chapitre

Dans ce chapitre, il a été question du choix des SOC à intégrer au S3M, des méthodologies d'intégration et de mise à jour de ces SOC mais également de leurs enrichissements. L'intégration d'un nouveau SOC impose un certain nombre d'étapes que sont la modélisation, la sélection puis le traitement des données. Il s'agit ensuite de contrôler l'intégrité des informations, de les insérer dans la BDD puis d'effectuer des processus de batchs et de validation. Nous avons vu, via plusieurs exemples, quelle était la démarche générale d'intégration de nouveaux SOC dans le S3M et comment il était possible d'enrichir leurs contenus et de construire un réseau sémantique permettant, entre autres, l'intéropérabilité. Les modèles et données stockés dans le S3M sont exploités par des applications à des fins de consultation, recherche, édition, etc. L'un des objectifs de ces travaux est en effet le développement d'une plateforme 3M adaptée aux humains et aux machines.

Chapitre 5

Mise en œuvre et exploitation des intégrations

Sommaire

Dans ce chapitre, nous présenterons l'outil central de ce projet, servant donc de démonstrateur aux travaux exposés dans ce mémoire. Nous détaillerons ensuite les objectifs de l'outil, les méthodes et technologies employées pour le développer. En outre, d'autres applications du S3M au sein de divers projets seront présentées.

5.1 Le Portail Terminologique de Santé

Au début de ces travaux de thèse, le Portail Terminologique de Santé (PTS) s'appuyait sur un modèle multi-terminologique bilingue français-anglais. Cette application est le fruit de l'implémentation de certaines méthodes issues du projet InterSTIS. L'outil avait pour but initial de rechercher des concepts dans plusieurs SOC simultanément pour aider les documentalistes dans leur tâche d'indexation des ressources CISMeF. La fonction principale de PTS était donc de trouver, parmi tous les termes des SOC sélectionnés, un concept pertinent correspondant à leur recherche. Les fonctionnalités sont présentées plus en détails dans la section dédiée à HeTOP, l'outil qui a pris la suite de PTS lors de l'implémentation du S3M.

5.2 HeTOP

HeTOP (pour Health Terminology/Ontology Portal : http://www.hetop.org/) est considéré comme la suite du PTS, en proposant de nouveaux SOC, l'univers inter- et multi-lingue et différentes améliorations et nouvelles fonctionnalités. Il s'agit donc de l'outil central développé pendant ces travaux de thèse. HeTOP, comme PTS, est une « vitrine » sur les SOC intégrés au S3M. Il permet donc de rechercher, de consulter et de naviguer entre les concepts des SOC mais aussi entre les SOC eux-mêmes. HeTOP propose un site web et un Service Web, accessibles partout depuis l'Internet et proposant des accès libres ou par authentification aux différents SOC. Le site web est dédié aux humains, avec un effort particulier sur la mise en page et la mise à disposition de données pertinentes à l'utilisation des SOC.
Toutes les fonctionnalités du PTS ont été reprises ainsi que son aspect graphique. La couche d'interface graphique est donc quasiment identique au PTS (même si le framework graphique a été changé et a donc nécessité un tout nouveau code informatique). Côté couches métier et données, il a fallu par contre tout modifier. HeTOP s'appuie sur des API de haut et bas niveaux développées par l'équipe technique CISMeF. En outre, un effort particulier a été apporté à l'optimisation des performances, autant sur les temps de réponses du moteur de recherche que sur la réactivité des composants de l'application.

5.2.1 L'approche orientée utilisateur

L'objectif principal de CISMeF est de créer des outils performants, pertinents, utilisables et compréhensibles pour ses utilisateurs. Dans le contexte de HeTOP, beaucoup d'éléments sont difficiles à appréhender par les non spécialistes du domaine de l'ingénierie des connaissances. Il en va de la même manière pour chaque SOC. Leurs vocabulaires et leurs fonctionnements spécifiques ne sont pas simples à présenter à l'utilisateur dans une interface conviviale. C'est la raison pour laquelle l'équipe s'attache a créer des interfaces graphiques simples mais complètes permettant d'exploiter au maximum les différents SOC. Un effort particulier est également apporté pour traduire les méta-données complexes en français (exemple des relations de la FMA) et pour proposer des raccourcis de qualité (accès aux ressources, etc.).

5.2.2 L'approche orientée machine

Durant la dernière année de cette thèse, un travail spécifique à l'exploitation du S3M par les machines a été effectué via la conception et le développement d'un Service Web.
Un Service Web (SW) est un programme permettant d'échanger des données sur le web. La plupart des SW répondent à des requêtes par l'intermédiaire de signatures, méthodes élémentaires proposées par les services. Ces signatures doivent être strictement définies (arguments et types d'entrée et de sortie, etc.) via une grammaire spécifique : le WSDL (*Web Services Description Language*). Les réponses renvoyées par les SW sont principalement structurées en XML ou en JSON [1]. Ces formats permettent de s'abstraire des langages de programmation informatique et d'appliquer des normes de standardisation. Par ailleurs, la plupart des SW communiquent les informations via le protocole HTTP [2] et définissent également des technologies pour encapsuler les données qui transitent : on peut citer le SOAP [3] et le REST [4], qui sont aujourd'hui les plus utilisées.
Dans l'objectif d'industrialisation et de mise à disposition de SOC, un SW HeTOP constitue un atout technologique de poids. Il est d'ailleurs aujourd'hui utilisé dans le cadre de deux ANR TecSan : SYNODOS (cf. 5.4.3) et SIFADO [5] afin d'accéder au contenu et aux alignements de certains SOC, et ce, dans des contextes distincts.

1. *JavaScript Object Notation*, format de données textuelles calqué sur la définition des objets du langage JavaScript : http://json.org/
2. *Hypertext Transfer Protocol*, protocole de couche de transport utilisé pour faire communiquer un serveur et un client : http://www.w3.org/Protocols/
3. *Simple Object Access Protocol*, protocole client-serveur bâti sur le XML permettant l'échange de messages normés sur un réseau : http://www.w3.org/2002/07/soap-translation/soap12-part0.html
4. *REpresentational State Transfer*, style d'architecture client-serveur servant à représenter un service via le web
5. Le projet SIFADO (pour Saisie Informatique FAcile de DOnnées médicales) vise à concevoir, développer et évaluer de nouvelles méthodes et outils ergonomiques pour faciliter la saisie et le codage de données dans les DPI : ANR-11-TECS-0014 , http://sifado.smbh.univ-paris13.fr/

Le SW HeTOP a été développé en Java grâce à l'API CXF [6] qui offre des facilités de déploiement en SOAP et en REST [Balani & Hathi, 2009].

Voici les différents fonctionnalités (sous-sections 5.2.3 à 5.2.9) conçues et développées pour HeTOP. Elles sont toutes disponibles via l'interface graphique et via le Service Web dédié, à part pour l'accès aux ressources, uniquement disponible via le site web.

5.2.3 Recherche 3M

HeTOP permet avant tout de rechercher des concepts terminologiques des SOC intégrés au S3M. Pour cela, l'utilisateur doit sélectionner des SOC, choisir une langue et entrer une expression. Tous les attributs terminologiques ne permettent pas de retrouver des concepts : en effet, afin de minimiser le bruit, d'augmenter la précision et ne pas diminuer les performances (temps de réponse), seules certaines méta-données ont été sélectionnées spécifiquement pour représenter les concepts. Il s'agit essentiellement des libellés préférés et des synonymes.
La recherche s'effectue sur la langue sélectionnée ainsi qu'en anglais, pour augmenter le rappel.
Par défaut, la recherche s'effectue avec des troncatures [7] à gauche et à droite de la requête ; cela permet également d'augmenter le rappel en s'affranchissant de certaines formes grammaticales (pluriel, ponctuation, etc.) et de l'ordre des mots. Il est bien sûr possible de désactiver cette option pour limiter un éventuel bruit. Par exemple, la requête « plastie » avec troncature (donc « *plastie* ») dans le MeSH renvoie 113 résultats (dont « arthroplastie », « galvanoplastie », etc.) alors que cette même requête sans troncature (juste le mot « plastie ») ne renvoie que 6 résultats (juin 2014).

5.2.4 Consultation d'un concept terminologique

Trois onglets servent à décrire complètement un concept de SOC dans HeTOP (trois signatures dans le SW). Ils correspondent respectivement en fait aux trois entités du méta-modèle du S3M et aux trois tables du modèle logique de données générique : `DATATYPE_PROPERTY` (Figure 5.1), `OBJECT_PROPERTY` (Figure 5.4) et `HIERARCHY` (Figures 5.2 et 5.3).

Description

La partie « Description » d'un concept terminologique concerne essentiellement tous ses attributs : libellé préféré, identifiant d'origine, CUI UMLS, définitions, syno-

6. http://cxf.apache.org/
7. Caractères joker ou *wildcard* en anglais

nymes, etc. Ces méta-données sont listées dans un ordre bien défini (par les experts), pour chaque type de concept de SOC. Le contenu est affiché dans les différentes langues disponibles. Sur HeTOP, il faut cliquer sur un bouton dédié pour faire apparaître toutes les langues.
Les icônes VCM (cf. 4.6.4) sont affichées, si elles existent, pour le concept sélectionné.
Enfin, des liens contextuels vers d'autres sites sont proposés dans un encart particulier.

FIGURE 5.1 – Capture d'écran de HeTOP : onglet Description du Descripteur MeSH « asthme »

Hiérarchies

Pour les concepts terminologiques impliqués dans des relations hiérarchiques, l'onglet « Hiérarchies » de HeTOP permet de visualiser l'emplacement du concept au sein d'un arbre. Il est possible de naviguer dans cet arbre en cliquant sur d'autres concepts ou en déployant les sous-arborescences via le signe « + ». Plusieurs hiérarchies peuvent être affichées les unes en dessous des autres si un concept est impliqué dans plusieurs hiérarchies différentes. Par ailleurs, comme expliqué dans la partie

4.7.2, les polyhiérarchies sont gérées dans l'affichage de HeTOP. Ainsi, un concept peut apparaître plusieurs fois dans la même hiérarchie, y compris avec le même BT (concept-père) mais avec des chemins différents. Enfin, une option permet de passer d'un mode de hiérarchie simple vers un mode de hiérarchie complète. La hiérarchie simple (par défaut) n'affiche que le concept terminologique ciblé (Figure 5.2), ses concepts-fils (NT) et tous ses ancêtres. La hiérarchie complète affiche en plus tous les concepts-frères du concept ciblé et de ses ancêtres (« concepts-oncles ») (Figure 5.3).

FIGURE 5.2 – Capture d'écran de HeTOP : onglet Hiérarchie simplifiée du Descripteur MeSH « asthme »

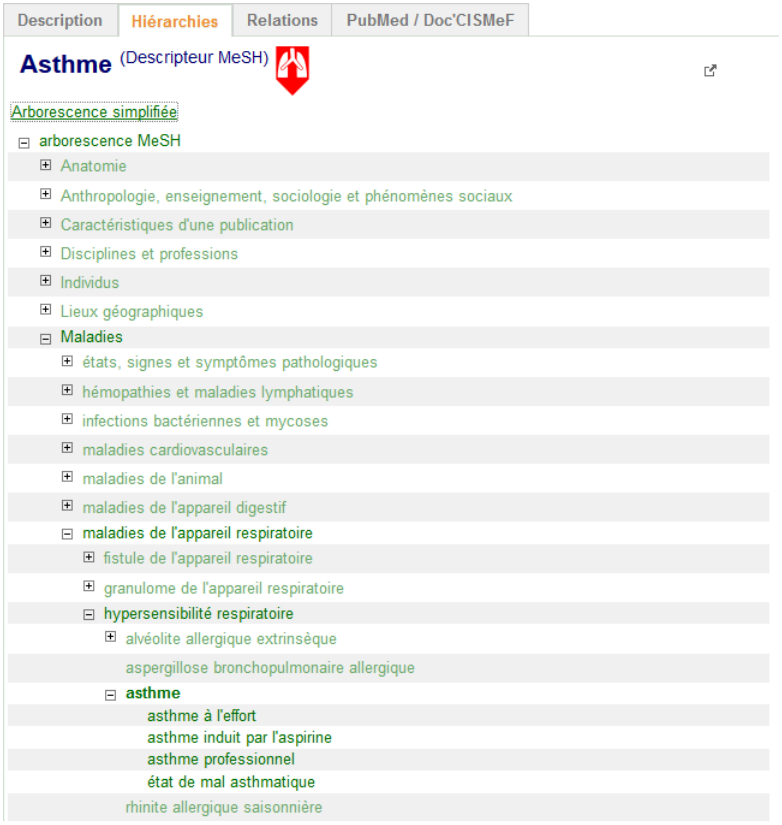

FIGURE 5.3 – Capture d'écran de HeTOP : onglet Hiérarchie complète du Descripteur MeSH « asthme »

Relations

L'onglet « Relations » affiche toutes les relations non hiérarchiques impliquant un concept terminologique. Celles-ci sont classées en deux catégories : relations intra-terminologiques et inter-terminologiques (Figure 5.4). Cet onglet peut s'avérer complexe à appréhender pour les non-initiés (cf. 6.4.3).

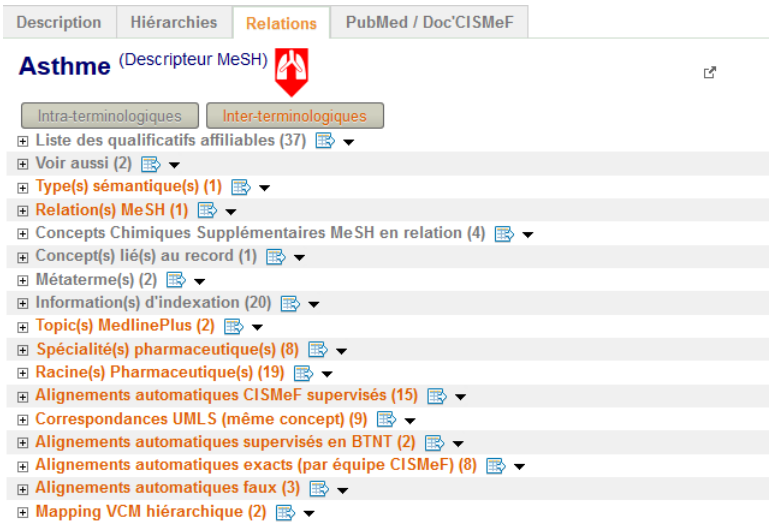

FIGURE 5.4 – Capture d'écran de HeTOP : onglet Relations du Descripteur MeSH « asthme »

Accès aux ressources

Le quatrième onglet de HeTOP « PubMed / Doc'CISMeF » permet d'accéder directement aux ressources MEDLINE et CISMeF indexées avec le concept sélectionné. Une série d'options est disponible pour construire une requête étendue (qualificatifs pour le MeSH, filtres sur les types de documents, etc.) (Figure 5.5). Cette requête va être automatiquement construite et envoyée au programme InfoRoute [8] qui s'occupera de l'interpréter pour l'étendre encore en fonction des spécificités des moteurs PubMed et Doc'CISMeF (cf. 5.2.6).

8. http://inforoute.chu-rouen.fr

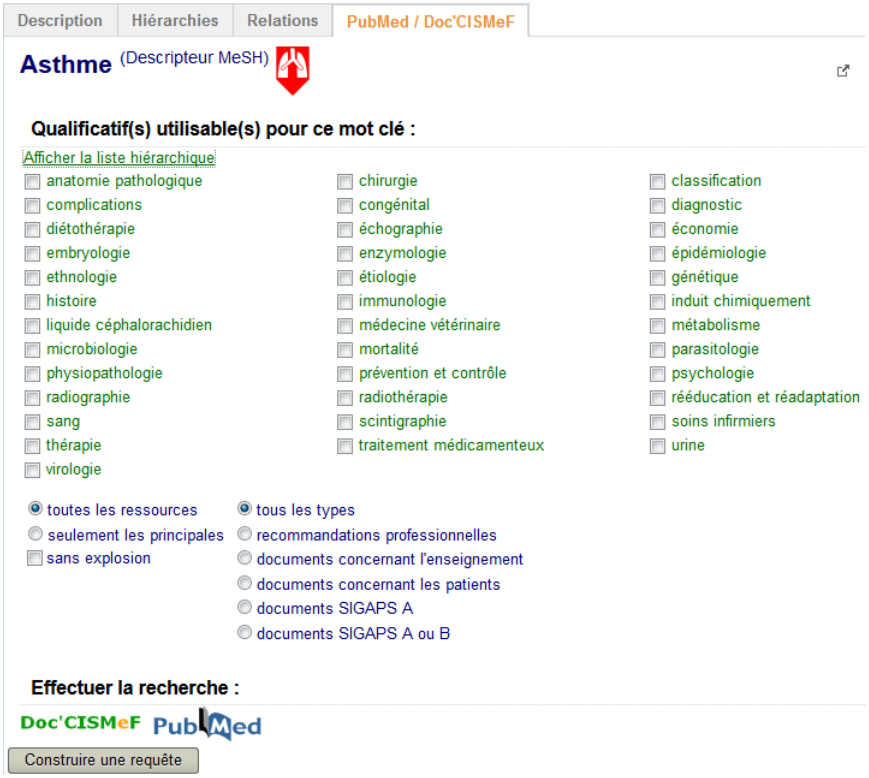

FIGURE 5.5 – Capture d'écran de HeTOP : onglet d'accès aux ressources pour le Descripteur MeSH « asthme »

5.2.5 Affichage des langues

La sélection et la priorité d'affichage des langues posent un certain nombre de problèmes d'utilisabilité. Dans HeTOP, on distingue 3 types d'éléments multi-lingues représentés différemment dans le SI :

- les libellés préférés des concepts terminologiques ;
- les libellés des méta-données (noms des SOC, types de concepts, types de relations, etc.) ;
- les libellés d'interface graphique (menus, boutons).

Pour simplifier l'utilisation de HeTOP, le choix de la langue est unique. Cette langue sera utilisée pour la RI, l'interface et les données. Si un libellé n'est pas disponible pour la langue sélectionnée, le programme affichera la valeur en anglais et sinon, une autre valeur possible. Ce mécanisme peut générer des mélanges de langues dans l'interface, ce qui peut apporter une certaine confusion. Pour le français et l'anglais, la plupart des libellés sont disponibles. Le nombre de données multi-lingues des méta-données et des éléments d'interface étant relativement faible (quelques milliers), certains partenaires scientifiques de l'équipe CISMeF à l'international ont pu traduire un bon nombre de libellés dans diverses langues (italien, norvégien, grec, arabe, etc.).

L'avantage du modèle logique de données générique est par ailleurs bien perceptible dans ce cadre. Il n'est pas nécessaire de changer le code informatique de HeTOP lors d'ajouts de traductions de libellés car tout est géré dans la BDD et dynamiquement dans les applications.

5.2.6 Accès multi-lingue à MEDLINE

MEDLINE (abrégé de *Medical Literature Analysis and Retrieval System Online*) est une base de données bibliographiques de la littérature scientifique et plus particulièrement en médecine, santé et biologie. Gérée et maintenue par la NLM, MEDLINE est le corpus le plus volumineux (21 millions de références en 2014 entre 1946 et aujourd'hui) et le plus utilisé de la communauté biomédicale. Les articles sont indexés manuellement par des spécialistes grâce au thésaurus MeSH. La plupart des références sont en anglais (92,4% en mai 2014) mais 60 autres langues sont tout de même disponibles.

PubMed [9] est le moteur de recherche couplé à MEDLINE est permet d'effectuer des requêtes simples ou complexes. Il est alors possible d'exploiter les spécificités du MeSH afin d'obtenir des résultats les plus précis possibles. Cependant, ces requêtes doivent être rédigées en anglais et reposent sur une grammaire complexe à appréhender.

L'une des fonctions de HeTOP est de permettre l'accès à MEDLINE dans un grand

9. http://www.ncbi.nlm.nih.gov/pubmed/

nombre de langues (toutes les langues du MeSH disponibles : français, portugais, tchèque, allemand, finlandais, italien, norvégien, russe, chinois, suédois, espagnol et hongrois). Pour cela, il faut sélectionner la langue désirée puis effectuer la recherche dans cette langue. Avec l'onglet « PubMed / Doc'CISMeF », il suffit alors de cliquer sur le bouton dédié pour accéder à PubMed avec une requête déjà rédigée en anglais, comprenant les opérateurs adéquats et une éventuelle expansion sémantique favorisant le rappel et la précision (cf. 5.4.4).
La communauté des utilisateurs médecins de HeTOP plébiscite cette fonctionnalité en utilisant d'ailleurs un outil dédié appelé « Constructeur de requêtes ». Il s'agit d'une petite fenêtre flottante intégrée à HeTOP permettant de récupérer des concepts terminologiques au fur et à mesure des recherches afin de créer une requête booléenne complexe (Figure 5.6). Une étude démarrera d'ailleurs fin 2014, en collaboration avec l'Hôpital Italien de Buenos Aeres, pour évaluer cet outil en anglais, espagnol et français.

FIGURE 5.6 – Capture d'écran de HeTOP : exemple d'utilisation du « Constructeur de requêtes »

5.2.7 Accès restreint

Étant donné la criticité de certains SOC et de certaines données mais aussi l'importance relative d'autres SOC (terminologies d'interface) vis à vis de la majorité des utilisateurs, un système d'authentification a été mis en place pour HeTOP. Chaque utilisateur possède un compte dans le SI de CISMeF pour avoir accès à l'application. Des droits sont octroyés par défaut à tous les utilisateurs et parfois, au cas par cas, des SOC particuliers sont ouverts. Ainsi, il est possible que certains SOC ne soient disponibles que pour certains utilisateurs. Par ailleurs, il existe un utilisateur par

défaut pour que l'application soit ouverte à tous sans avoir à s'authentifier.
Il est à noter qu'il est possible de visualiser certains concepts de SOC dont on ne possède pas les droits via l'onglet des relations. En effet, si un concept (dont on a l'accès) possède des relations terminologiques vers des concepts de SOC non autorisés, il sera impossible de cliquer dessus pour avoir plus d'informations : seul son libellé préféré est visible.

5.2.8 Proposition de contenu par les utilisateurs

Depuis début 2013, une fonctionnalité permet aux utilisateurs inscrits de HeTOP de proposer leur propre contenu au S3M. Via un bouton dédié pour chaque concept terminologique, ils peuvent soit proposer une traduction, soit un nouveau synonyme. Ces propositions sont ensuite validées ou invalidées par les experts de l'équipe CISMeF. Dans le cas d'une validation, le synonyme est donc ajouté dans le S3M (avec un type particulier) et sera affiché dans HeTOP et pourra même servir pour la RI. Seuls certains types de concepts sont concernés mais l'utilisation de cette fonctionnalité reste aujourd'hui limitée en terme de nombre de propositions (6.4.2).

5.2.9 Catalogue des SOC disponibles

La liste des SOC disponibles pour un utilisateur donné est accessible soit dans la barre de sélection des SOC, soit via l'onglet « Terminologies » de la page d'accueil. Un certain nombre de méta-données sont disponibles pour chaque SOC : nom complet, éditeur, version, langue(s) originale(s), langues disponibles, description et d'éventuels liens directs pour consulter la ou les hiérarchie(s) disponible(s) pour chaque SOC.

5.3 Méthodes et outils de développement

HeTOP a été développé selon une architecture 3-tiers classique. Ce type d'architecture est défini comme un système divisé en trois couches :

- la couche présentation (« couche haute »), qui correspond aux interfaces graphiques en interaction directe avec les utilisateurs ;
- la couche métier, qui correspond à l'ensemble des programmes et règles qui traitent les données manipulées ;
- la couche de données (« couche basse »), qui correspond au sous-système responsable du stockage et de l'accès aux données persistantes (SGBD).

La Figure 5.7, détaillée dans la section suivante, illustre cette architecture souple et robuste, permettant notamment de changer de technologies dans une couche sans impacter les autres.
Concrètement, le S3M constitue la couche de données de HeTOP. Ce dernier pos-

sède une couche métier constituée de parties spécifiques mais également des API génériques couplées à la BDD (DBCore). Côté présentation, HeTOP s'appuie sur un framework Java puissant appelé Vaadin [10]. Il s'agit d'un outil de la catégorie des *Rich Internet Applications* (RIA) qui permettent aux développeurs web de s'abstraire des codes HTML, CSS [11] et JavaScript nécessaires à la création de pages web. Il est alors possible de lier directement les couches présentation et métier dans un ensemble de classes Java. Vaadin propose en fait des composants graphiques natifs (appelés « widgets ») à insérer directement dans des pages. La génération des codes HTML, CSS et JavaScript est automatique. Ce type de framework permet non seulement de faciliter le développement et la maintenance de l'application mais permet en outre de profiter de la puissance de Java et d'assurer la portabilité des sites web.
Malgré ses avantages indéniables, Vaadin ne constitue pas la solution idéale puisqu'il est peu adapté à la création de sites web à haute disponibilité et à la navigation via les URL. Son atout réside plutôt dans la création d'applications web « statiques », du type applications « locales de bureau ». Cela pose d'ailleurs un problème d'utilisabilité identifié après plusieurs semaines d'utilisation de HeTOP(cf. 6.4.3).

5.3.1 Environnements

Après plusieurs années d'utilisation de PTS, lorsque HeTOP a vu le jour, l'une des clauses du cahier des charges était de créer un site web modulable ; en effet, plusieurs utilisations se sont dégagées au fil du temps, dont celle de l'enseignement, du codage ou encore de l'accès à PubMed en français. Ces différentes variantes vont proposer des éléments graphiques et des options par défaut différents. Pour faciliter les développements, la maintenance et la modularité, nous avons décidé d'introduire dans le site web de HeTOP la notion d'« environnements ». Un environnement HeTOP est donc une variante du portail, souvent dédiée à une utilisation précise. L'environnement par défaut (« basic ») est celui de base dans HeTOP. Voici la liste détaillée des autres environnements en activité (disponibles via le menu d'options) :

- « PubMed » : dédié à l'accès à PubMed dans une langue autre que l'anglais. L'onglet d'accès aux ressources est montré en premier ;
- SFMU (pour Société Française de Médecine d'Urgence [12]) : version du portail dédiée au codage du DPI dans le contexte du service des Urgences au CHU de Rouen. L'outil du SIH permettant le codage n'est pas jugé suffisant par les utilisateurs. Pour palier ce problème, le souhait a été de créer cet environnement centré sur le thésaurus de la SFMU et la CIM-10. L'objectif est de rendre satisfaisant la RI dans ces SOC dans un minimum de temps pour aider

10. https://vaadin.com
11. *Cascading Style Sheets* ou « feuilles de style en cascade » est un langage informatique servant à la mise en page de documents HTML ou XML
12. www.sfmu.org

les cliniciens au codage dans un cadre opérationnel ;
- SIFADO (projet ANR mentionné précédemment) est l'environnement permettant également l'aide au codage, cette fois dans le contexte de la recherche (démonstrateur).

5.4 Une plateforme 3M pour de multiples applications

L'un des objectifs de cette thèse est de montrer l'importance d'un système unique supportant la multi-terminologie, le multi-discipline et le multi-linguisme car il offre un tronc solide à un SI reposant sur des SOC. L'exploitation d'une telle plateforme permet de bénéficier des dernières mises à jour des SOC, d'un réseau sémantique riche, de l'interopérabilité entre de nombreux SOC et enfin d'une base de connaissances extrêmement variée et complète.
Dans le cas du SI de CISMeF, cette plateforme 3M est un pivot essentiel autour duquel la plupart des applications gravitent car s'appuient toutes sur un ou plusieurs SOC (qu'il s'agisse d'un catalogue de ressources ou d'un système de gestion de Dossiers Patients Informatisés).

5.4.1 Le SI du CISMeF

Le SI de l'équipe CISMeF possède une architecture 3-tiers et repose sur une couche de données persistantes gérée par le SBGD Oracle (11g r2). Le S3M est une partie de cette BDD puisqu'il s'agit d'un modèle générique de données intégrant à la fois les données terminologiques mais aussi d'autres données utiles aux autres applications (catalogue CISMeF, etc.). La couche donnée est en fait doublée par une sous-couche de « cache ». Un système de cache permet de stocker des informations en mémoire vive afin d'accélérer ses accès.
La couche métier du SI CISMeF est divisée en deux sous-couches : (i) la partie « composants métier » proprement dite, qui contient les programmes permettant d'accéder à la BDD et procédant à différents traitements ; (ii) la partie « Services », qui présentent et utilisent les différentes applications via des SW ou d'autres programmes.
Enfin, la couche présentation correspond aux différentes interfaces des outils, comme HeTOP par exemple.
Tous ces éléments sont représentés dans la Figure 5.7. On peut notamment y voir le composant « parsers » qui désigne les P1 et P2 détaillés dans le chapitre 5, qui permettent l'intégration et la mise à jour des SOC.
Par ailleurs, toutes les applications présentées ici reposent sur le S3M. Nous détaillerons ci-après différentes applications au S3M utilisées quotidiennement.

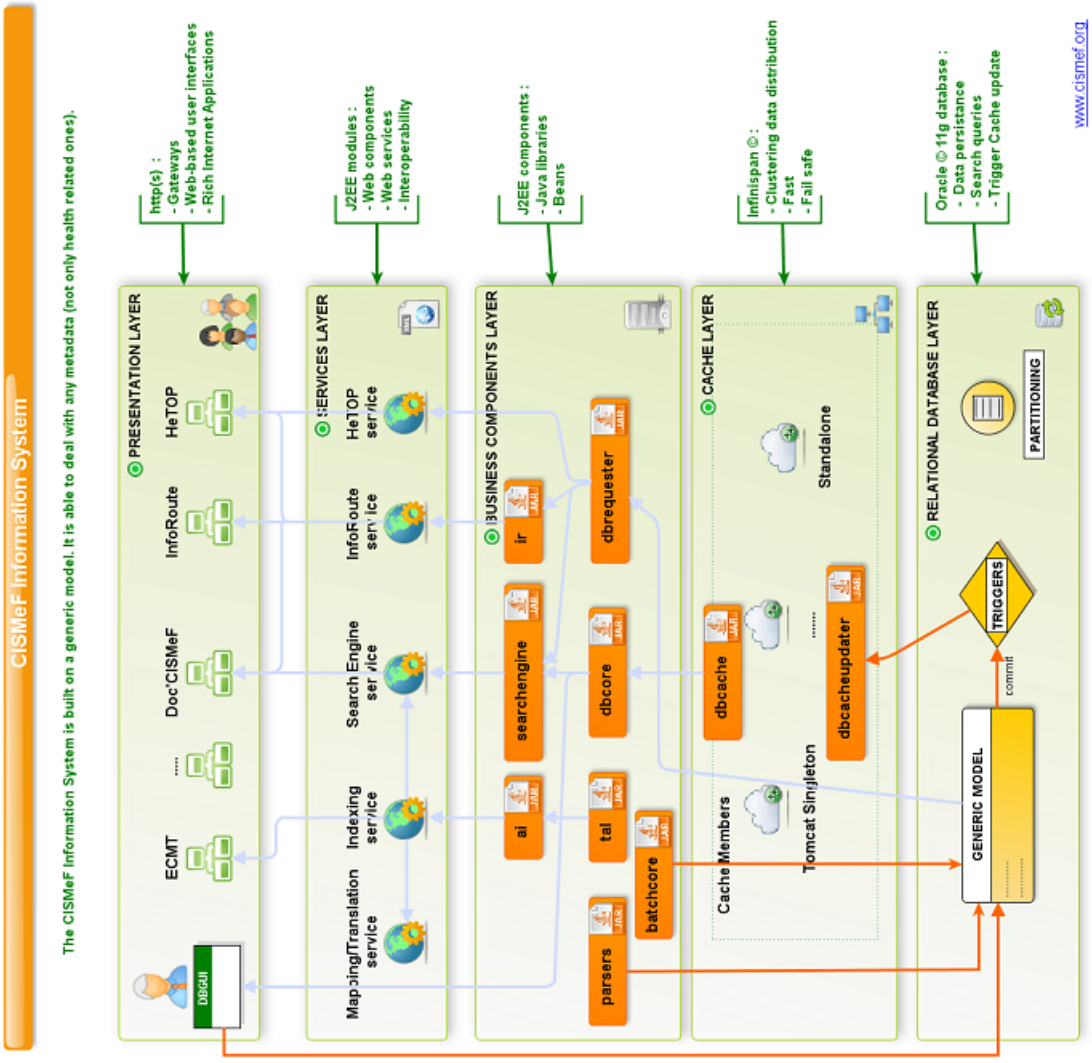

FIGURE 5.7 – Schéma détaillé du SI du CISMeF au mois de juin 2014

5.4.2 Le S3M pour la traduction et les alignements exacts

Les algorithmes d'alignements exacts automatiques présentés en 4.6.1 sont non seulement utilisés en routine lors d'intégrations ou de mises à jour de SOC dans le S3M mais ils peuvent également servir pour des utilisateurs voulant aligner ou traduire des termes. En effet, les algorithmes d'alignements se basent sur la similarité entre une requête et les termes (PT et synonymes) de concepts des SOC. Si des concepts de SOC correspondent bien à cette requête, non seulement cela constitue un alignement exact potentiel mais les différents libellés des termes dans plusieurs langues (exceptée celle de la requête) sont autant de traductions possibles à cette requête.

Ainsi, l'outil MT@HeTOP (pour Mapping/Translation via HeTOP) est une application web centrée sur cet algorithme. Encore en version de test [13], l'outil permet donc

13. http://cispro.chu-rouen.fr/MT_EHTOP/

d'aligner exactement une requête vers une sélection de SOC ou bien de la traduire. Les résultats apparaissent sous forme de tableaux avec les détails des termes trouvés (codes, libellés, types de concepts, SOC). La Figure 5.8 montre un exemple d'alignement exact avec la requête « foie » en français ; 5 termes sont alignés correspondant à 4 concepts de 4 SOC différents.

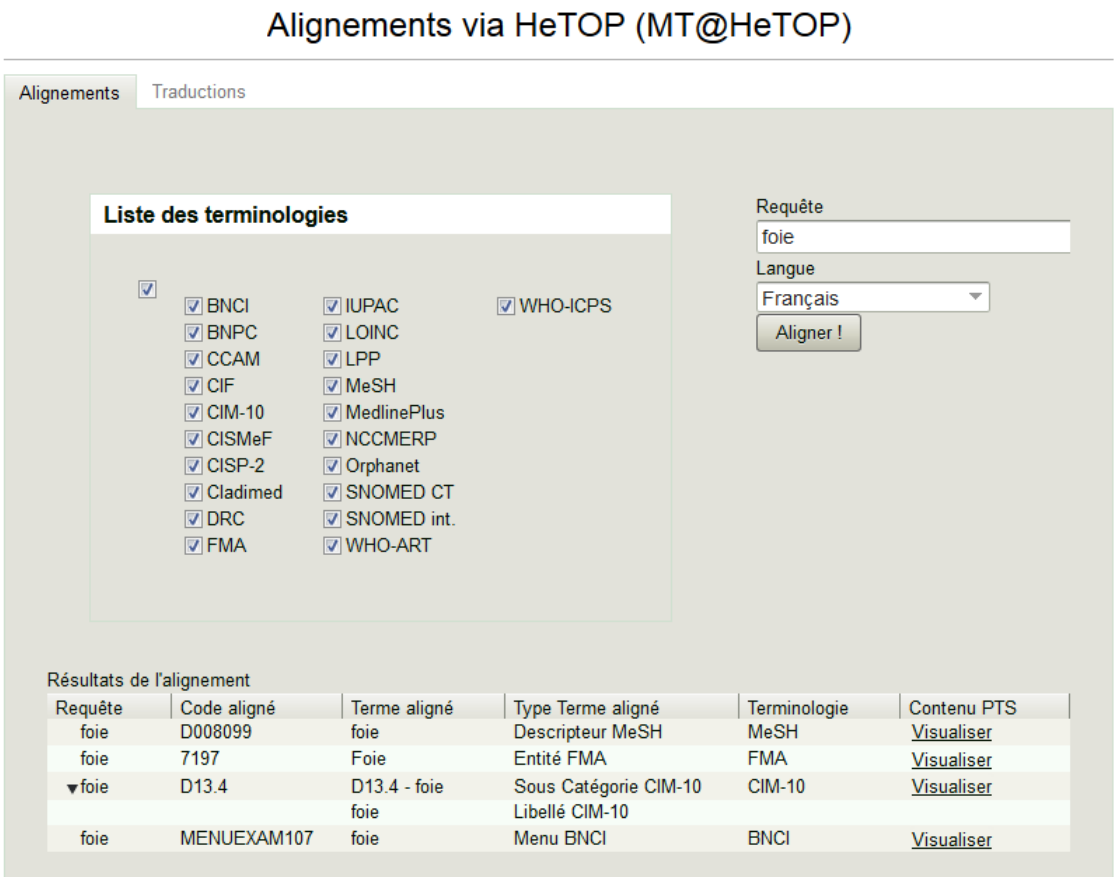

FIGURE 5.8 – Capture d'écran de l'outil MT@HeTOP : alignement exact d'un terme

5.4.3 Le S3M pour l'indexation

Indexation manuelle

Le S3M est initialement prévu pour indexer les documents du catalogue CISMeF. À ce titre, il est quotidiennement utilisé par les documentalistes de l'équipe via l'outil DBGUI (cf. 5.5). L'application leur permet de sélectionner des concepts via une méthode d'auto-complétion et de les rattacher à un document. La démarche classique d'un expert consiste d'abord à rechercher des concepts précis et pertinents via HeTOP pour ensuite les utiliser pour indexer un document.
Les concepts terminologiques du S3M sont également utilisés pour l'indexation et le codage dans d'autres domaines documentaires au sein du SI CISMeF : (i) intégration

des indexations des articles MEDLINE en français (projet déposé BDBfr [14]), (ii) intégration des codes PMSI des Dossiers Patient Informatisés (projets RIDoPI [15] et RAVEL [16]).

Annotation automatique

Que ce soit pour le catalogue CISMeF ou pour d'autres applications, un outil d'annotation automatique a été développé par l'équipe il y a quelques années. En effet, la plupart des documents disponibles sur l'Internet, dans le DPI ou dans bien d'autres domaines ne sont pas indexés nativement. Il est souvent difficile, à partir d'un texte, de déterminer automatiquement les mots-clés pertinents et représentatifs du document. Il s'agit d'une problématique extrêmement étudiée.
Dans un contexte de RI et de documents orientés Santé, l'équipe CISMeF a donc conçu un outil dédié à la tâche d'indexation automatique, dérivé des travaux de [Pereira *et al.*, 2008]. Ce programme, appelé Extracteur de Concepts Multi-Terminologique (ECMT), permet d'identifier des concepts de SOC dans des textes, principalement en français. Pour cela, des algorithmes de TAL sont appliqués et la méthode du « sac de mots » est utilisée.
L'ECMT repose alors sur le S3M pour rechercher les concepts. Il est utilisé quotidiennement en batch pour l'indexation de certains documents CISMeF. Par ailleurs, il est décliné en SW, notamment pour les besoins du projet SYNODOS [17] [Gicquel *et al.*, 2014].

5.4.4 Le S3M pour la RI

Le S3M est naturellement utilisé pour la RI documentaire. En effet, les documents étant indexés avec les concepts terminologiques, il faut, à partir d'une requête en langage naturel ou d'une requête logique, retrouver les documents concernés.

14. Le projet BDBfr a pour but de proposer une Base de Données Bibliographique en français via un moteur de recherche capable de fonctionnalités sémantiques. Le site proposera en outre des services à forte valeur ajoutée autour de la sélection et de la bibliométrie
15. Le projet RIDoPI (Recherche d'Information dans le Dossier Patient Informatisé) est un projet interne à l'équipe CISMeF et au CHU de Rouen visant à créer un modèle de données compact et pertinent afin d'améliorer la consultation des dossiers et de permettre une RI avancée : création de cohortes de patients, indicateurs de qualité, etc.
16. RAVEL (pour *Retrieval And Visualization in ELectronic health records*) est un projet financé par l'ANR qui consiste en la mise en pratique de RIDoPI pour des cas d'usages précis avec l'aide de technologies de pointe en TAL et en RI : ANR-11-TECS-0012
17. Le projet ANR SYNODOS (SYstème de Normalisation et d'Organisation de Données médicales textuelles pour l'Observation en Santé) vise à améliorer le soin et les études épidémiologiques en analysant les données textuelles de comptes-rendus médicaux : ANR-12-TECS-0006, http://www.synodos.fr/

Moteur de recherche

Le moteur de recherche CISMeF (CSE pour *CISMeF Search Engine*) est constitué de deux sous-programmes : (i) le premier se charge d'identifier les concepts terminologiques correspondant à la requête, (ii) le second construit et exécute une requête SQL afin de récupérer dans la BDD les documents indexés avec les concepts trouvés.

Des études ont d'ores et déjà montré l'intérêt d'une indexation et d'un RI multi-terminologiques, notamment au sein de Doc'CISMeF [Sakji, 2010], [Soualmia *et al.*, 2013]. La Figure 5.9 illustre la RI et l'indexation multi-terminologique dans Doc'CISMeF : ici, le document trouvé est indexé avec 26 concepts appartenant à 7 SOC différents.

FIGURE 5.9 – Capture d'écran de Doc'CISMeF : document indexé et retrouvé en multi-terminologie

InfoRoute

InfoRoute [18] est une application de la famille des « infobuttons », c'est-à-dire un programme permettant d'accéder rapidement à d'autres sites et portails à partir d'un mot-clé. InfoRoute propose plus de 50 liens directs vers des moteurs de recherche de qualité en français et anglais. Ceux-ci sont répartis en 12 catégories en fonction de leurs cibles (étudiants, patients, médicaments, maladies rares, essais cliniques, etc.). L'un des liens les plus utilisés et le plus complexe est celui vers PubMed. Celui-ci étant basé sur le MeSH, un travail spécifique a été effectué afin d'élaborer une requête très pertinente en s'aidant : (i) des méta-données et opérateurs MEDLINE (MH, TW, SB, etc.), (ii) du S3M qui permet de faire une expansion de requête via le réseau sémantique. Cette expansion consiste en l'ajout de synonymes MeSH et d'autres termes via les alignements exacts validés manuellement, mais pas via les alignements conceptuels de l'UMLS jugés trop bruyants. Ainsi, si un utilisateur entre la requête « acrodermatite » en français, InfoRoute reconnaîtra le Descripteur MeSH correspondant (D000169) et pourra construire une requête complexe augmentant le rappel, sans pour autant amener du bruit [Darmoni *et al.*, 2008]. Ici, on obtient 2 342 résultats avec la requête générée automatiquement (cf. Figure 5.10) alors que la requête simple « `acrodermatitis »[MH]` ne renvoie que 1 926 résultats.
InfoRoute est directement intégré à HeTOP via un SW dédié et permet l'accès à PubMed et à Doc'CISMeF via l'onglet d'accès aux ressources.

```
("acrodermatitis"[MH] OR ("acrodermatitis"[TW] OR "papular
acrodermatitis, infantile"[TW] OR "acrodermatitis papulosa
infantum"[TW] OR "gianotti crosti syndrome"[TW] OR "acropapulo
vesicular syndrome"[TW] OR "Erythemato-Vesiculo-Papulous eruptive
syndromes"[TW] OR "Gianotti-Crosti syndrome"[TW] OR "syndromes,
Acropapulo-Vesicular"[TW] OR "Acropapulo-Vesicular syndromes"[TW]
OR "acrodermatitis papulosa infantums"[TW] OR "infantile papular
acrodermatitides"[TW] OR "papulovesicular acrolocated syndromes"[TW]
OR "erythemato vesiculo papulous eruptive syndrome"[TW] OR
"papulovesicular acrolocated syndrome"[TW] OR "papular acrodermatitis
of childhood"[TW] OR "infantile papular acrodermatitis"[TW]
OR "Erythemato-Vesiculo-Papulous eruptive syndrome"[TW] OR
"acrodermatitis, infantile papular"[TW] OR "acrodermatitides, infantile
papular"[TW]))
```

FIGURE 5.10 – Exemple d'expansion sémantique de InfoRoute pour le Descripteur MeSH « acrodermatitis »

18. http://inforoute.chu-rouen.fr/

5.5 Gestion et édition des SOC : l'outil DBGUI

Tout SI documentaire doit intégrer un outil de saisie afin de créer, de modifier ou de supprimer des informations. Dans le cas du CISMeF, l'équipe technique a choisi de refondre totalement son outil de saisie lors de la migration au niveau modèle de base de données présenté dans ce mémoire. Cela a permis en outre d'ajouter de nouvelles fonctionnalités, de mettre l'interface graphique au goût du jour et d'utiliser des technologies web plus avancées afin d'améliorer les capacités de l'outil.
Le DBGUI (pour *DataBase Graphic User Interface*) a donc été conçu et développé avec plusieurs objectifs principaux :

- Proposer un outil au moins aussi performant et fonctionnel que l'ancien (non régression) ;
- Améliorer les fonctionnalités et ajouter des composants adaptés aux utilisateurs, notamment pour la gestion des SOC ;
- Construire une « Rich Internet Application » (RIA) avec des outils web novateurs ;
- Concevoir cet outil non pas pour la seule problématique du CISMeF mais comme un outil générique couplé à la BDD, étant elle-même générique .

Cette application a, comme HeTOP, été développée grâce au framework Vaadin, qui dans cas, montre toutes ses capacités.
Dans l'objectif de ces travaux de thèse, le DBGUI devait donc être l'outil d'édition des SOC, autant pour la partie des modèles que pour celles des concepts.

5.5.1 Gestion des modèles conceptuels (édition partie MODEL)

Le DBGUI intègre plusieurs formulaires permettant de créer et d'éditer des TI. Cela facilite bien sûr la saisie dans la BDD mais assure surtout le contrôle des contraintes liées à certaines méta-données. En effet, une fenêtre d'édition de TI permet de visualiser et d'éditer ses différentes propriétés (MDP, MOP, MIN) mais affiche également certains champs spécifiques à remplir obligatoirement en fonction du TD du TI (contraintes). La Figure 5.11 est une capture d'écran d'une création de TI avec les différentes méta-données à remplir. Les MDP, MOP et MIN sont représentés en tableaux, tels qu'ils le sont réellement dans la BDD. L'héritage des TI est interprété puis affiché dans les catégories de propriétés correspondantes (Figure 5.12).
D'autres formulaires permettent de définir l'ordre des méta-données à afficher dans les applications ou même d'ajouter des entrées dans la BDD à la volée.

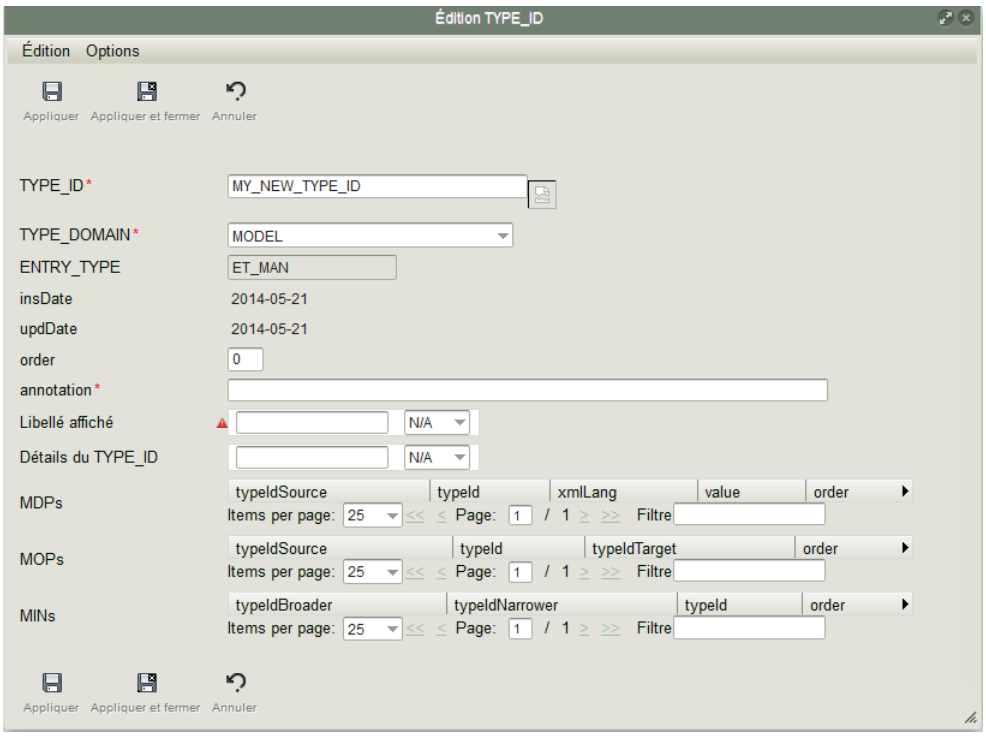

FIGURE 5.11 – Capture d'écran du module de création de TYPE_ID dans DBGUI

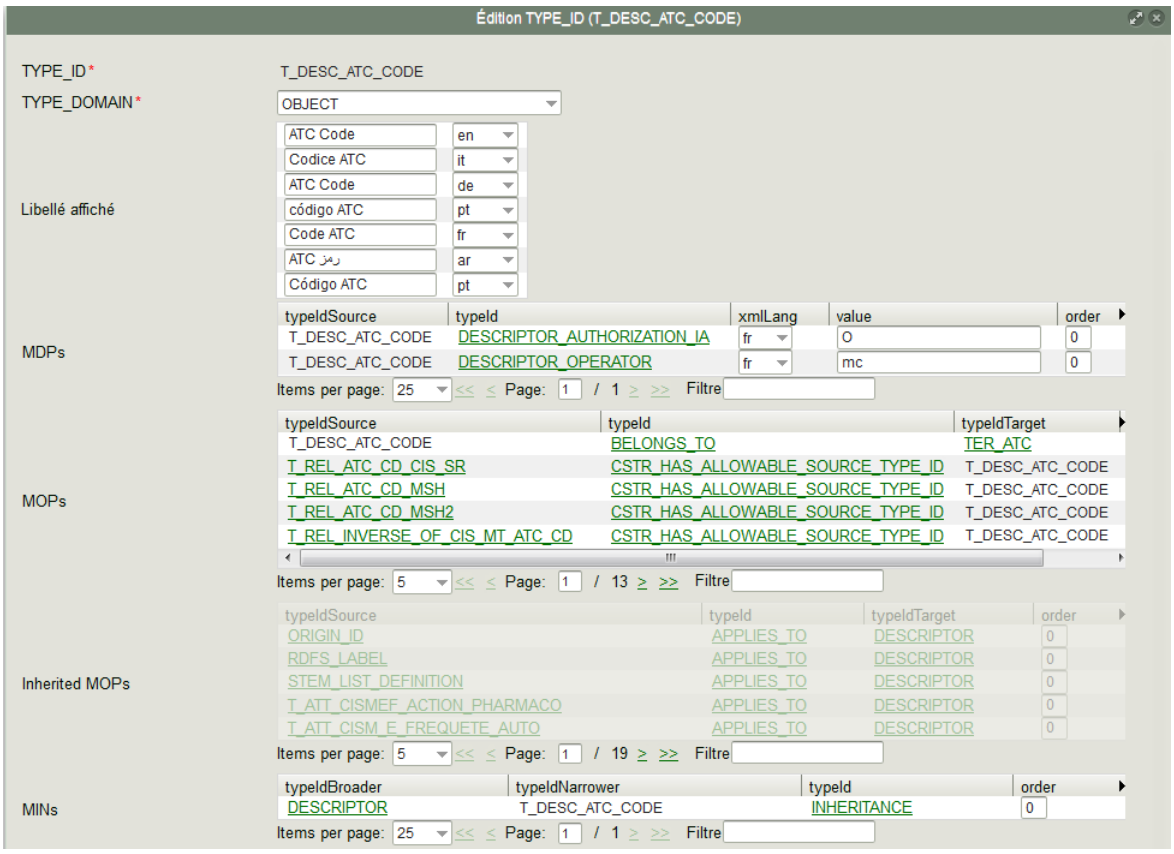

FIGURE 5.12 – Capture d'écran du module d'édition de TYPE_ID dans DBGUI : Code ATC

Ces interfaces sont exclusivement réservées aux administrateurs du SI CISMeF pour éviter toute erreur. Les différentes contraintes et règles s'appliquant à la partie MODEL de la BDD peuvent avoir un impact immédiat sur les applications. Il est donc nécessaire d'avoir une expertise du système pour effectuer toute modification.

5.5.2 Gestion des objets (édition partie OBJECT)

La partie la plus importante du DBGUI pour les utilisateurs est la gestion des objets ; c'est-à-dire l'édition des ressources stockées dans la BDD. Pour cela, il existe deux modules principaux : (i) affichage en liste par TI (Figure 5.13) et (ii) fiche d'édition de ressource. Celle-ci est divisée en 4 parties au maximum, comme montré sur la Figure 5.14 ; ces 4 parties correspondent aux 4 entités du modèle générique de données : les attributs (DATATYPE_PROPERTY), les relations hiérarchiques (HIERARCHY), les relations d'indexation (INDEXING) et les autres relations (OBJECT_PROPERTY). Des options sont disponibles pour afficher telle ou telle langue, cacher certaines méta-données par défaut ou encore annoter des propriétés.
Ces fonctionnalités sont génériques puisqu'elles concernent tous les types d'objets de la BDD. En effet, ces modules se basent sur la partie MODEL qui contient les différentes méta-données à afficher dans un ordre précis. Les codes informatiques de l'application ne sont donc pas à modifier pour l'ajout de nouveaux SOC ou de nouvelles propriétés.
Le DBGUI est utilisé quotidiennement par les experts de l'équipe CISMeF éditant les SOC. Les opérations les plus fréquentes sont : traduction de terme, ajout de synonymes, validation d'alignements automatiques et ajout de d'alignements manuels.

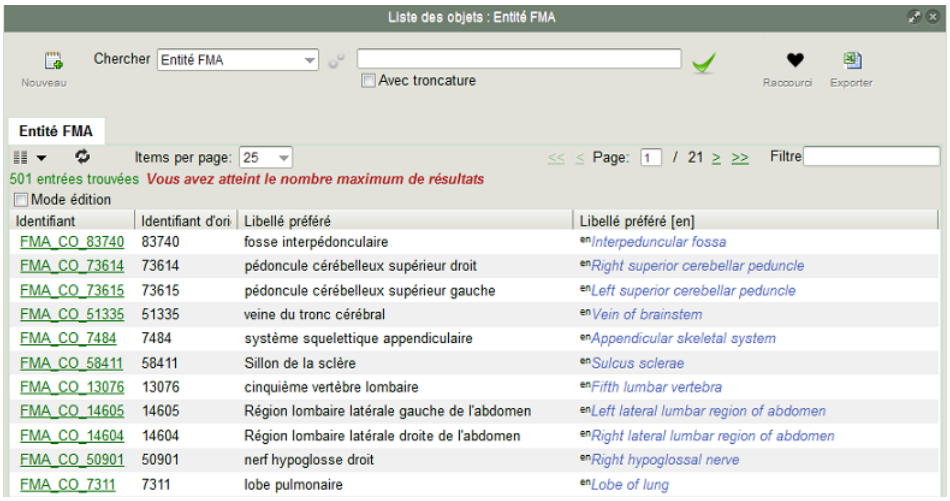

FIGURE 5.13 – Capture d'écran du DBGUI : affichage en liste des Entités FMA

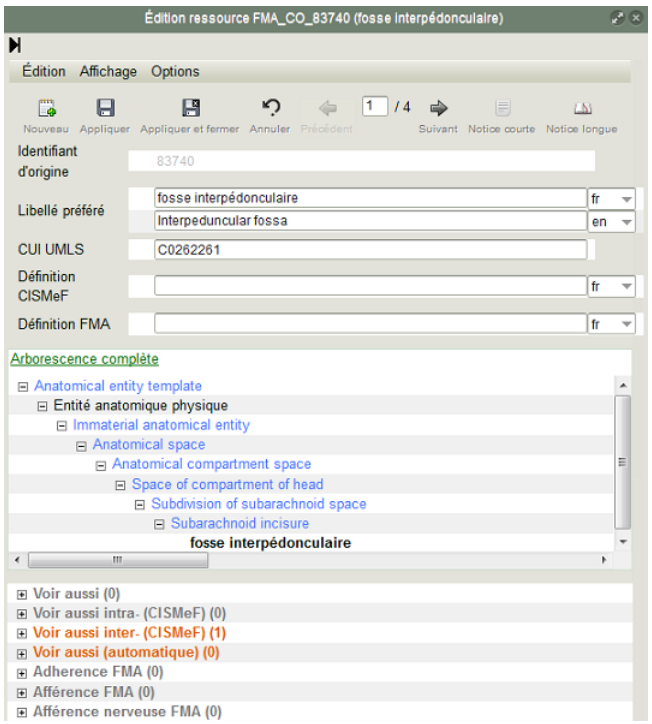

FIGURE 5.14 – Capture d'écran du DBGUI : édition de l'Entité FMA « fosse interpédonculaire »

5.5.3 Outils dédiés

Étant donné qu'à chaque nouvelle intégration ou mise à jour de SOC dans le S3M, des processus sont exécutés automatiquement pour aligner ou traduire les concepts (cf. 4.7), la conception d'outils spécialisés dans la validation facilitée de ces traitements a vite été nécessaire. Les experts ont donc émis certains souhaits quant aux fonctionnalités de base et deux outils ont vu le jour pour les aider dans leurs tâches de validations. Dans les deux cas, il s'agit de tableaux interactifs où il est possible de consulter les concepts terminologiques en cliquant sur leur identifiant. Il est également possible d'ajouter des annotations particulières pour commenter les alignements ou les traductions.

Outil de validation des alignements

Le DBGUI intègre donc un module spécifique à la validation des alignements exacts automatiques. Il est également possible d'ajouter directement des traductions via ce module lorsque l'on valide un alignement. L'exemple montré en Figure 5.15 concerne la validation d'alignements de OMIM et sa traduction en français. Si l'expert estime que l'alignement proposé est bien exact, il coche l'option « Alignement automatique CISMeF supervisé » ; si un des deux termes alignés possède un libellé préféré français et que l'autre n'en possède pas, le libellé français est automatiquement proposé pour la traduction du second.
Les experts peuvent néanmoins choisir d'autres types d'alignements s'ils ne consi-

dèrent pas l'alignement comme exact. Les autres choix sont les suivants : alignements plus larges (BTNT), plus précis (NTBT), relié (voir-aussi) ou bien faux (invalidation). Il est parfois très difficile de qualifier un alignement ; dans ce cas, il est possible de cocher une option « ? » pour spécifier que l'alignement ne peut pas être encore qualifié. Cet alignement sera alors toujours considéré comme automatique mais sera placé en fin de liste des éléments à valider.

FIGURE 5.15 – Capture d'écran du module de validation d'alignements de DBGUI

Outil de traduction

Un autre module est dédié spécialement à la traduction. Pour cela, on choisit un type de concept à traduire et une langue. La liste est automatiquement générée dans un tableau avec des formulaires pour ajouter la traduction en face du libellé préféré d'origine. La Figure 5.16 montre un exemple de liste de concepts HPO à traduire en français à partir de l'anglais.

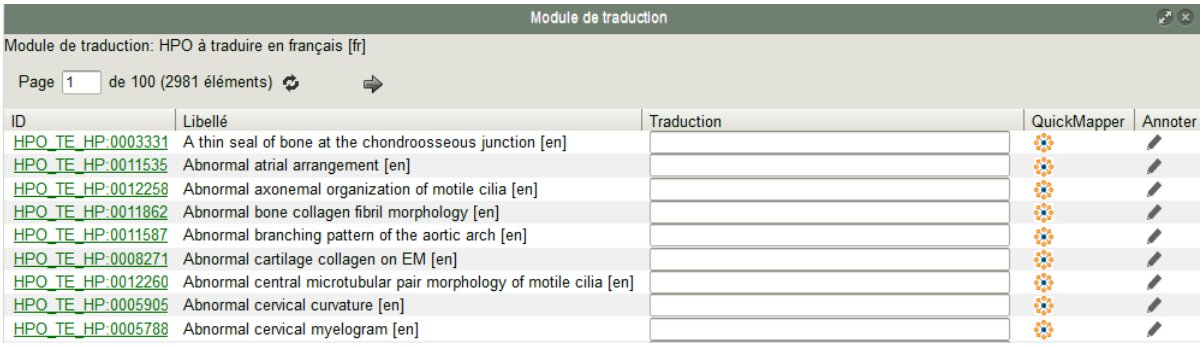

FIGURE 5.16 – Capture d'écran du module de traduction de DBGUI

5.5.4 Exemple de la création d'un SOC

Dans le cadre du projet ANR SYNODOS, une terminologie a été créée. En effet, certains termes sont trouvés fréquemment dans des comptes-rendus médicaux mais n'existent pas dans les principaux SOC de Santé en français. Afin de pouvoir retrouver ces termes et les documents les contenant, l'équipe CISMeF a été chargée de créer une terminologie dédiée contenant des termes considérés comme importants (en tous cas, dans le cadre du projet). Ainsi, la « terminologie SYNODOS » est divisée en deux types de concepts (voir le modèle en Figure 5.17) :

1. les Combinaisons SYNODOS qui sont des libellés français rattachés à des concepts d'autres SOC ;
2. les Termes SYNODOS qui sont aussi des libellés français mais rattachés à aucun autre concept.

Lorsque ces concepts sont rencontrés lors de l'indexation automatique (par l'ECMT, cf. 5.4.3), le document est indexé avec ceux-ci ainsi qu'avec les concepts d'autres SOC reliés dans le cas des Combinaisons.
Les différents TI nécessaires à cette terminologie ont d'abord été créés puis édités via le DBGUI.
Désormais, l'expert en charge de la création des termes s'aide des fiches d'édition pour créer et éditer les concepts. Les processus de normalisation se font en temps réel et les libellés sont directement utilisables par l'ECMT pour l'indexation automatique.

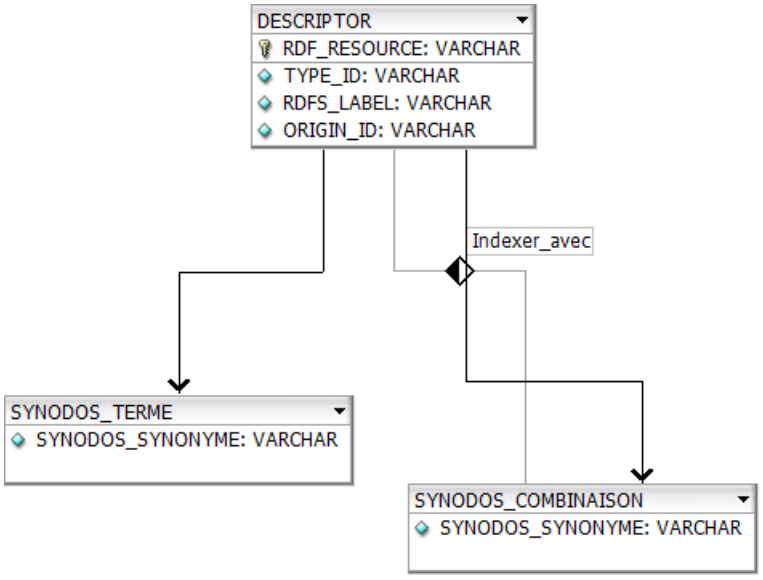

FIGURE 5.17 – Modèle terminologique de la Terminologie SYNODOS au sein du S3M

5.6 Synthèse du chapitre

Dans ce chapitre, deux applications ont été présentées. D'une part, HeTOP est le portail permettant de rechercher des concepts dans un environnement 3M. Il facilite la navigation, au sein d'un SOC via les hiérarchies et les relations intra-terminologiques, mais il propose également de parcourir les SOC via le réseau sémantique. D'autres fonctionnalités sont offertes, autant aux humains qu'aux machines, pour utiliser ces SOC (accès à d'autres portails, à des moteurs de recherche bibliographiques, accès multi-lingue, etc.).
D'autre part, le DBGUI a été développé pour l'édition des ressources stockées dans le modèle générique de données. Nous avons ensuite vu que les outils de RIA permettaient le développement facilité de modules dédiées à la gestion des modèles et des données des SOC. Cet outil est utilisé quotidiennement par les experts de l'équipe CISMeF pour ajouter ou valider du contenu.
L'objectif majeur du S3M est de constituer une plateforme à de nombreux usages. Nous allons voir comment elle est utilisée aujourd'hui. Enfin, pour valider et consolider les outils créés, il est nécessaire de les évaluer via différents indicateurs et via la comparaison à des systèmes proches.

Chapitre 6

Résultats, évaluations et applications au SI de CISMeF

Sommaire

Dans ce chapitre, je présenterai les résultats des intégrations de SOC dans le S3M et de HeTOP. Il est question d'évaluation de ces méthodes, aussi bien qualitatives que quantitatives. En outre, je discuterai des apports de ces travaux pour la communauté et le positionnement de nos outils par rapport aux systèmes proches existants.

6.1 Validation des intégrations

Assez rapidement, le problème de la validation des intégrations s'est posé. Comment être sûr d'avoir bien toutes les données aux bons endroits (bonnes métadonnées) et bien structurées ? Or, il n'existe pas, à notre connaissance, de méthode

spécifique reconnue permettant de valider ou d'évaluer la qualité d'intégration de données d'un système à un autre.
Dans notre étude, la première méthode de validation se fait via le programme Outils SMTS (cf. 4.4.4) puisqu'il est possible de contrôler les volumétries, les concepts orphelins, etc. Ceci dit, cette étape s'effectue sur le fichier intermédiaire de données et non pas dans le S3M directement. Il a donc fallu écrire un certain nombre de requêtes SQL pour contrôler l'intégrité des données insérées. Ce sont pour la plupart des étapes de comptage que l'on compare avec les chiffres initiaux. Une méthode complémentaire, systématiquement appliquée, consiste à sélectionner un échantillon de données à partir des sources du SOC et de les confronter avec leurs équivalents dans le S3M via des requêtes SQL (ou directement dans HeTOP).
Malgré ces contrôles, il existe encore régulièrement des incohérences de données, soit parce-que la source initiale comporte des erreurs ou des cas particuliers ou soit parce que certaines subtilités ou anomalies métiers apparaissent lors d'une phase de l'intégration. On peut notamment citer des soucis d'encodage multiple ou caractères spéciaux, des valeurs aberrantes d'attributs (vides, par exemple), etc. Dans ces cas, il est nécessaire d'effectuer des vérifications et des réparations spécifiques manuellement. Heureusement, ces évènements restent très rares.

6.2 Désavantages d'un serveur 3M

Certains inconvénients à l'utilisation du S3M ont été découverts au fur et à mesure des intégrations et exploitations. Cela n'avait pas été anticipé pour plusieurs raisons.
Premièrement, l'utilisation à grande échelle du S3M dans des contextes d'indexation/annotation automatiques de d'expressions en langage naturel (ou textes) apporte un grand nombre de problématiques linguistique, sémantique, techniques, etc. De plus, les exploitations du réseau sémantique se sont faites au fur et à mesure et le nombre de SOC intégrés augmentait. De ce fait, chaque nouvelle intégration peut apporter un lot de biais, d'erreurs et d'approximations pouvant affecter le reste du système.

6.2.1 Perte de contexte

Dans un serveur multi-terminologique et multi-discipline, l'un des plus grands dangers est d'utiliser les termes hors de leur contexte, c'est-à-dire hors de leur SOC d'origine. Intégrer beaucoup de SOC dans un seul système ne peut pas être une solution basique consistant à créer un énorme sac de termes. Il n'y a aucun intérêt à utiliser les concepts isolément. Comme expliqué plusieurs fois dans ces travaux, les concepts terminologiques n'ont de valeur que lorsqu'ils sont utilisés dans leur contexte et dans l'objectif de chaque SOC. Il faut donc faire des efforts pour rassem-

bler les SOC de mêmes buts et séparer ceux qui sont trop différents afin de fournir aux utilisateurs un catalogue dans lequel ils peuvent piocher mais en gardant à l'idée que les concepts manipulés ont une histoire et une portée spécifique.

6.2.2 Portée des erreurs

Comme mentionné plusieurs fois dans cette étude, les éventuelles erreurs sur les libellés (pertes de contexte, problèmes de traduction) et sur les alignements peuvent avoir des conséquences importantes sur d'autres concepts. L'établissement d'un réseau sémantique possède des avantages indéniables mais s'avère parfois être un véritable danger quant à la véracité des informations. Ainsi, il arrive relativement fréquemment que des experts découvrent des erreurs sur des éléments qu'ils n'ont pas saisi manuellement. Cela est essentiellement dû au fait que les traitements automatiques ou semi-automatiques se basent sur le réseau sémantique pour entrer certaines informations.
Il est très difficile de quantifier ces types d'erreur et il est parfois nécessaire d'effectuer des traitements spécifiques pour « réparer » les données.
Un exemple est celui des synonymes hérités via les alignements conceptuels. Ces alignements sont issus de l'UMLS et constituent le cœur de ce méta-thésaurus. Cela dit, ils ne sont pas exempts d'erreurs et cela engendre des non-sens importants. Par exemple, le Concept SNOMED CT « *disorder characterized by eosinophilia (disorder)* » (419455006) est aligné conceptuellement avec le Descripteur MeSH « éosinophilie » (D004802) car le Concept SNOMED CT possède ce terme en synonyme. Cela est incohérent, les deux concepts ne sont pas équivalents *stricto sensu*.
Il faut donc être extrêmement vigilant lors des modifications pour ne pas impacter des parties entières du réseau sémantique. On observe parfois un effet « boule de neige » qu'il est difficile de corriger.

6.3 Erreurs et autres problèmes natifs des SOC

Dans les deux chapitres précédents, j'ai notamment expliqué comment et pourquoi l'exploitation des SOC et de leur interopérabilité était importante. Traduire, enrichir et aligner les SOC sont autant de challenges complexes auxquels le serveur 3M essaie de répondre au mieux. Cependant, le travail que j'ai effectué sur la modélisation mais aussi les différents processus automatiques, semi-automatiques et manuels d'enrichissement des SOC ont permis de repérer un bon nombre de sources d'erreurs.
Au niveau modélisation, la redondance de concepts à des niveaux distincts a été un souci pour le MeSH, MedDRA et WHO-ART. Il a fallu adapter leurs modèles en tenant compte de leurs spécificités.
Au niveau du contexte des termes, plusieurs axes de la CIM-10 possèdent des libellés

inappropriés. Il a fallu les exclure de certains traitements.
Dans la gestion de la synonymie, beaucoup de SOC introduisent des termes en synonymes de concepts mais ils relèvent souvent d'hyponymes ou d'hyperonymes. Aucun traitement automatique fiable n'a pu être mis en place pour ce genre de problème.
Il existe également des termes polysémiques (orthographes identiques mais sens différents). Il est possible de les repérer mais un traitement manuel spécifique doit être fait pour les qualifier.

Quelques solutions

Outre les problèmes de modélisation de SOC qu'il a fallu résoudre en les adaptant légèrement, la majorité des problèmes encore récurrents aujourd'hui concerne la qualité et la représentativité des termes par rapport à leur concept. En plus des erreurs de traductions et de synonymie, s'ajoutent des problèmes de contextes et d'utilisation ; l'ajout de synonymes est très subjectif. Qu'il s'agisse d'acronymes, d'abréviations, de dérivés, de raccourcis ou d'autres diverses subtilités orthographiques et syntaxiques, les synonymes (au sens large) constituent une source de problèmes importante pour la connaissance elle-même mais surtout pour la RI. Une solution méthodologique est en cours d'élaboration (cf. 7).

6.4 Bilans et évaluations du S3M

Bien évidemment, les nombres présentés ci-après évoluent quotidiennement, compte-tenu des SOC intégrés ou mis à jour régulièrement. Plusieurs métriques permettent de faire un bilan analysable du S3M. De la même façon, via les logs de requêtes et de trafic du site HeTOP, il est possible de tirer quelques conclusions quant à son utilisation et son utilisabilité. Je ferai donc ici les bilans sur les intégrations de SOC dans le S3M, mais également sur leurs enrichissements. Par la suite, je tirerai quelques conclusions sur HeTOP via différentes statistiques et les avis des utilisateurs.

6.4.1 Bilan des SOC intégrés

Le Tableau 6.1 dresse un bilan chiffré des intégrations des SOC dans le S3M.
Au total, 58 SOC ont été intégrés. Ceux-ci sont principalement issus du domaine de la Santé (n=54). Le côté multi-discipline a été exploré, au sens de la preuve de concept, via l'intégration de SOC d'autres domaines comme l'IDIT [1], pour les Droits du Transport ou encore l'UNIT [2] pour les sciences de l'ingénieur. En effet, dans le cadre du projet PlaIR, ces deux terminologies ont fait l'objet d'analyses afin d'étudier la compatibilité de leurs modèles avec le méta-modèle 3M. Celui-ci a prouvé sa généricité et ces deux SOC ont été intégrés avec succès et sont consultables via HeTOP.
Le S3M compte plus de 2 200 000 concepts. Plus de 50% d'entre eux sont présents dans seulement 3 SOC : MeSH (26,8%), SNOMED CT (13,4%) et CIM-10 PCS [3] (10%).
Les concepts sont représentés par plus de 7 500 000 termes, soit un ratio de 3,4 termes par concept terminologique. Ceci s'explique non seulement par le fait que les PT sont souvent multi-lingues mais aussi par le fait que beaucoup de synonymes existent.
Enfin, si l'on additionne le nombre de concepts (dans TB_OBJECT), le nombre d'attributs (dans TB_DATATYPE_PROPERTY) et le nombre de relations (dans TB_OBJECT_PROPERTY et TB_HIERARCHY), cela donne plus de 23 950 000 de lignes dans la base de données.

1. L'Institut du Droit International des Transports a pour but de proposer une base de données de publications en Droit du Transport, d'organiser des manifestations et des formations mais également d'offrir une expertise dans ce domaine : http://www.idit.asso.fr/
2. L'Université Numérique Ingénierie et Technologie est une université numérique thématique autour des sciences de l'ingénieur. Elle propose des outils et des documents aux étudiants et enseignants : http://www.unit.eu/
3. La CIM-10 *Procedure Coding System* est un système de codage d'actes médicaux basé sur la génération combinatoire de termes à partir de primitives.

	Nombre total
SOC intégrés	58
Concepts	2 218 972
Termes (PT et synonymes)	7 568 181
Synonymes	3 792 925
Définitions	238 626
Autres attributs	13 155 802
Relations (dont hiérarchies)	8 581 559

TABLE 6.1 – Tableau récapitulatif des SOC intégrés au S3M

6.4.2 Bilan des enrichissements de SOC

Que ce soit avant ou pendant ces travaux, les SOC ont subi des ajouts permettant d'enrichir leurs contenus. Le Tableau 6.2 récapitule différents types d'ajouts dans le S3M, que ce soit manuellement ou automatiquement. Les relations correspondent aussi bien aux alignements (cf. 6.4.2) qu'aux différents traitements d'enrichissement sémantique (cf. 4.6.3). Certains SOC ont connu des traitements plus spécifiques avec l'ajouts des icônes VCM, de requêtes pré-définies pour des moteurs de recherche bibliographiques, etc. Il en va de même avec des campagnes de traductions en français comme avec HPO (89,2%), NCIT (37,6%) ou encore la SNOMED CT (36,1%). D'ailleurs, nous montrons avec Névéol [Névéol *et al.*, 2014] que l'apport à l'UMLS en terme de traduction française est non négligeable. Le Tableau 6.3 recense le nombre de concepts en français dans l'UMLS et dans le S3M ainsi que les sources de ces traductions : on peut dénombrer 105 891 concepts traduits en français dans l'UMLS contre 523 077 de plus dans le S3M, soit presque 5 fois plus (+494%).
Ces informations constituent un atout de poids pour le S3M car elles permettent non seulement de compléter et d'améliorer la connaissance mais elles améliorent surtout l'utilisation des SOC : RI, consultation, indexation, interopérabilité, etc.

	Nombre total
Traductions	294 043
Synonymes	100 593
Autres attributs	649 020
Relations	1 047 905

TABLE 6.2 – Tableau récapitulatif des enrichissements de SOC dans le S3M

SOC	Traductions françaises	Dans UMLS	Sources
ATC	5 834 (100%)	non	VIDAL et CISMeF
CIF	1 496 (100%)	non	CISMeF
CIM-9	10 716 (100%)	non	CISMeF
CIM-10	40 804 (100%)	non	ATIH
CIM-10 PCS	7 297 (5%)	non	CISMeF
CISP-2	746 (100%)	oui	CISP-2
FMA	10 265 (18,2%)	non	CISMeF et FMA
GO	557 (1,6%)	non	CISMeF
ICNP	2 812 (100%)	non	ICNP
LOINC	57 942 (60,1%)	non	AP-HP
MedDRA	74 411 (100%)	oui	MedDRA
MEDLINEplus	847 (99,8%)	non	LIMSI et CISMeF
MeSH	159 796 (26,8%)	partiellement	DISC et CISMeF
NCIT	31 628 (37,6%)	non	CISMeF
OMIM	6 570 (84,5%)	non	CISMeF
SNOMED CT	107 063 (36,1%)	non	CISMeF
SNOMED int.	106 266 (100%)	non	ASIP Santé
WHO-ART	3 483 (100%)	oui	WHO-ART
WHO-ICPS	435 (67%)	non	CISMeF

TABLE 6.3 – Tableau des traductions en français de concepts UMLS

Bilan des alignements, le réseau sémantique

Les processus d'alignements exacts automatiques TAL sont exécutés systématiquement à chaque nouvelle intégration ou mise à jour de SOC. Cela génère évidemment beaucoup de relations qu'il faut ensuite valider. Comme il s'agit d'un processus très couteux en temps et obligatoirement manuel, les experts font ce travail au fil de l'eau, en se concentrant sur certains SOC en priorité.
Les alignements conceptuels sont ceux issus de l'UMLS.
Le Tableau 6.4 récapitule les différents nombres d'alignements inclus dans le S3M. On peut voir qu'il reste encore beaucoup d'alignements automatiques exacts TAL à valider. Ceci dit, on peut dresser un bilan intéressant concernant la précision des algorithmes de TAL via la validation des experts de l'équipe : 89,7% des alignements sont considérés comme valides en *exact match*, 6,1% sont considérés comme plus larges ou plus étroits (*close match* en BT-NT ou NT-BT) et seulement 4,2% ont été étiquetés comme faux. Il faut ajouter que quasiment la moitié des alignements exacts validés (48,5%) l'ont été via les algorithmes de validation par transitivité (cf. 4.7.3). Enfin, une étude plus poussée serait nécessaire pour évaluer le rappel des algorithmes d'alignement.
Si l'on compte plus de 1 370 000 alignements exacts, cela ne correspond qu'à 597 134 concepts différents soit 26,9% des concepts du S3M. En effet, la plupart des SOC du S3M n'appartiennent pas à l'UMLS (donc pas d'alignements conceptuels directs) et beaucoup de SOC sont très spécifiques et possèdent des libellés difficilement comparables à d'autres sans traitements préalables (CCAM, LOINC, CIM-10 PCS, etc.). Si on soustrait les alignements exacts non supervisés, cela correspond à 19,8% des concepts du S3M impliqués dans le réseau sémantique.

Type d'alignement	**Nombre**
Exacts TAL automatiques	658 303
Exacts validés	247 470
Exacts manuels	37 337
Exacts conceptuels	430 821
Exacts total	1 373 931
BT-NT validés/manuels	1 825
NT-BT validés/manuels	14 940
Exacts TAL faux	11 533
Exacts conceptuels faux	397

TABLE 6.4 – Tableau récapitulatif des alignements du S3M

L'apport de la communauté

Durant la seconde moitié de ces travaux de thèse, un objectif a été de mettre à contribution les utilisateurs de HeTOP. Nous voulions les solliciter afin qu'ils nous aident sur le contenu des SOC en ajoutant des traductions des PT mais aussi des synonymes. Cela s'est fait essentiellement via le bouton dédié dans HeTOP (cf. 5.2.8) mais aussi récemment via une aide précieuse de plusieurs utilisateurs désireux d'aider la communauté.
Pour le moment, ces contributions sont encore marginales (67 traductions et 24 synonymes) mais nous espérons rendre ces mécanismes plus visibles et plus valorisants dans HeTOP dans les mois à venir.

6.4.3 Bilan d'utilisation de HeTOP

La première version fonctionnelle de HeTOP a vu le jour en juin 2012. Il s'agissait uniquement de la version site web, avec une vingtaine de SOC disponibles dans une dizaine de langues, uniquement latines. Les différentes évolutions techniques et les nouvelles intégrations ont été faites au fur et à mesure avec, par exemple, l'ajout de nouvelles langues non latines comme le japonais, l'arabe, etc. La première version de production de HeTOP a été mis en ligne en janvier 2013. À partir de janvier 2014, HeTOP a définitivement supplanté le PTS puisqu'il reprenait au moins toutes les fonctionnalités de ce dernier tout en supportant la charge de production.
L'outil AWSTATS[4] permet de tracer les accès des utilisateurs à des pages web. Ainsi, il est possible de suivre le trafic sur un site en nombre d'utilisateurs différents, visites, pages, hits, etc. Depuis sa mise en production « complète » (à la place du PTS), HeTOP est consulté par environ 7 500 visiteurs différents chaque mois, ce qui correspond à environ 600 visites par jour travaillé (et deux fois moins le week-end). 75% des requêtes viennent de France puis 6% de Belgique et 3% du Canada. D'autres pays, principalement francophones, viennent compléter ces statistiques : Suisse, Tunisie, Italie, États-Unis, Algérie, Liban, Maroc, ...

Les utilisateurs inscrits

HeTOP compte aujourd'hui 1 696 utilisateurs inscrits. Ceux-ci possèdent donc un compte personnel avec plus de SOC disponibles que le compte standard. Lors du processus d'inscription sur le site, plusieurs informations sont demandées, dont l'activité (profession ou études) et le pays. Le Tableau 6.5 référence les professions des utilisateurs inscrits à HeTOP et de façon logique, la majeure partie concerne des métiers de la Santé (37,9% au total) et l'enseignement (étudiants et enseignants : 42,1%). Ces résultats vont à l'encontre du but initial du PTS, qui visait un public de documentalistes/indexeurs. Même si le PTS a rapidement vu sa cible dérivée

4. http://awstats.sourceforge.net/

vers un outil de codage (médecins), l'utilisation principale est aujourd'hui celle de l'enseignement. HeTOP est aujourd'hui un support de communication pour les professionnels de Santé, les chercheurs et les formateurs qui enseignent la médecine, la pharmacie ou encore l'informatique médicale. La moitié des écoles de médecine en France mais aussi d'autres facultés, au Canada par exemple, possèdent des cours orientés sur HeTOP et Doc'CISMeF.
La grande majorité des utilisateurs résident en France (65%), puis viennent le Canada (7,1%) et la Belgique (3,5%). Beaucoup de pays d'Afrique francophones sont aussi représentés, même minoritairement (total de 13,7%). En tout, les utilisateurs de HeTOP travaillent dans 48 pays différents. Cela s'explique notamment par les différentes collaborations internationales (Corée du Sud, Japon, Grèce, etc.) et par la bonne visibilité des outils CISMeF depuis près de 20 ans.

Catégorie de profession	**Nombre total (%)**
Étudiants (internes, externes, élèves infirmiers, ...)	654 (38,5%)
Médecins (libéraux, PH, dentistes, etc.)	486 (28,6%)
Documentalistes/bibliothécaires	127 (7,5%)
Chercheurs/ingénieurs	85 (5%)
Infirmiers/kinésithérapeutes	66 (3,9%)
Traducteurs	64 (3,8%)
Enseignants/formateurs	61 (3,6%)
Pharmaciens	26 (1,5%)
Autres professionnels de santé	65 (3,8%)
Autres	62 (3,6%)

TABLE 6.5 – Tableau des principaux profils utilisateurs inscrits à HeTOP

Les requêtes

Un système de suivi des requêtes a été implémenté dans HeTOP. Chacune d'entre elles est entrée dans la base de données et plusieurs propriétés sont stockées : temps de réponse, nombre de réponses, langue de requêtes, environnement, etc.
En termes de temps de réponse, 98% des requêtes sont exécutées en dessous de 5 secondes et 79,1% en dessous de 1 seconde. Si l'on exclut les quelques points aberrants (essentiellement dûs à des anomalies métiers ou réseau ponctuelles), le temps de réponse moyen est de 0,81 seconde. Tout ceci va dans le bon sens pour l'utilisabilité de HeTOP puisqu'une limite a été fixée à 2 secondes maximum de temps de réponse en moyenne dans le cahier des charges initial de HeTOP. Il est également intéressant d'observer son évolution en fonction du nombre de SOC sélectionnés. La Figure 6.1 montre que cette évolution est linéaire (R^2=0,9823). Ceci est moins bon que prévu même si les temps de réponse moyens restent acceptables. Nous espérions en fait avoir un premier plateau puis un accroissement linéaire. En effet, les outils Oracle permettent en théorie d'effectuer des calculs en parallèle. Ceux-ci ne semblent pas fonctionner correctement malgré de nombreuses tentatives. Une autre explication est que le temps mesuré ne correspond pas uniquement au temps de recherche dans la base de données mais également à l'affichage des résultats. Plus il y a de concepts trouvés, plus cette dernière étape peut s'avérer longue (système de cache dynamique).
On peut observer 230 238 requêtes en l'espace de 121 jours de production soit environ 1 902 requêtes par jour (en comptant les week-end, beaucoup moins chargés en terme de trafic).
En ce qui concerne les SOC sélectionnés lors des recherches sur HeTOP, 81,2% des requêtes sont effectuées sur le MeSH et CISMeF qui sont les terminologies par défaut. Il est difficile de savoir si cette utilisation massive correspond au véritable besoin des utilisateurs pour ces vocabulaires ou bien s'il s'agit d'un biais constitué par le fait qu'il s'agisse des valeurs par défaut. De la même façon, 98,6% des requêtes sont effectuées en français/anglais.
Si le nombre moyen de réponses (n=42,8) n'est pas réellement analysable, il est plus intéressant de regarder le nombre de requêtes qui ne renvoient aucun résultat. En l'occurence, pas moins de 66 405 requêtes n'ont abouti à aucune réponse de la part de HeTOP, soit 28,8% du total des requêtes. Ce nombre est plus important que prévu. Une étude poussée n'a pas encore été menée pour comprendre l'origine de ce phénomène mais des premières estimations et explications sont avancées : (i) au moins 4% des requêtes à 0 réponse possèdent des problèmes d'encodage de caractères et le moteur ne les interprète pas bien ; (ii) 78,4% de ces requêtes ne sont faites que sur MeSH/CISMeF par défaut alors que d'autres SOC pourraient, eux, renvoyer des résultats ; (iii) un nombre non négligeable de requêtes mal orthographiées a également été décelé en analysant ces logs. Tout ces éléments expliquent, en

partie, ce nombre élevé de requêtes sans réponse de la part de HeTOP. Cependant, il est également normal que beaucoup de termes n'existent pas encore dans le S3M, tous les domaines et sous-spécialités ne sont pas encore représentés et tous les SOC sont loin d'être exhaustifs.
Sur les 1 690 utilisateurs inscrits, 40 d'entre eux seulement se connectent régulièrement avec leur compte personnel. Le compte par défaut de HeTOP (donc sans s'authentifier) est largement majoritaire (88,5% des requêtes). Pour l'instant, nous estimons que cette utilisation marginale des comptes utilisateurs personnels est essentiellement due à deux facteurs : (i) les SOC uniquement disponibles pour les inscrits ne sont pas assez intéressants, (ii) l'utilisation principale de HeTOP concerne le MeSH.
Nous espérons très bientôt mener une campagne d'analyse poussée des logs pour évaluer finement la RI dans HeTOP et ainsi essayer d'améliorer le système.

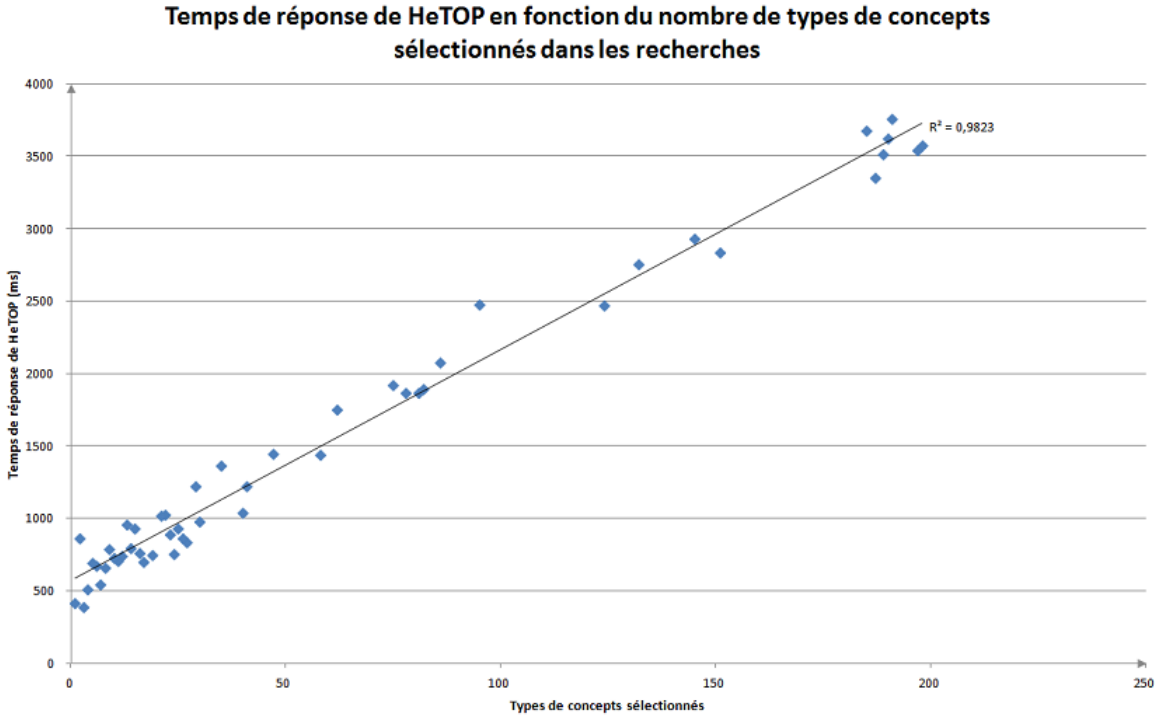

FIGURE 6.1 – Graphique de l'évolution des temps de réponse moyens dans HeTOP en fonction du nombre de SOC sélectionnés

Service Web

En deux mois d'activité, le Service Web HeTOP est surtout utilisé à des fins de recherche pour le projet SIFADO. Sa visibilité est encore volontairement peu étendue pour conclure une phase de tests et de validations. Nous espérons bientôt plus largement diffuser ce service pour les spécialistes même si son utilisation va probablement être limitée. En effet, notre expérience sur ce type d'outil tend à montrer que les Services Web sont encore peu utiles pour la plupart des utilisateurs

puisque leur intérêt principal consiste à faire communiquer des machines entre elles. Or, HeTOP n'a pour l'instant pas d'interactions avec d'autres systèmes. Il s'agit cependant d'un objectif à moyen terme.

Évaluation par les utilisateurs

Nous avons mis en place un questionnaire en ligne pour recueillir les impressions et les utilisations de HeTOP. Disponible depuis mai 2014[5], il est composé de 9 questions : 4 questions basées sur une échelle de Likert, 4 questions type QCM et 1 question libre.
Une échelle de Likert (du nom de son concepteur Rensis Likert) est une échelle de mesure psychométrique utilisée principalement dans des questionnaires. Le principe est de proposer 5 ou 7 degrés d'adhésion d'un utilisateur à une question. À chaque degré est associée une valeur numérique et l'ensemble est ainsi facilement analysable via des indicateurs de statistiques descriptives classiques (moyenne, écart-type). Cette méthode est désormais utilisée fréquemment et montre régulièrement son efficacité [Maurer & Pierce, 1998].
Jusqu'à maintenant, 64 réponses ont été collectées. Le Tableau 6.6 recense les différentes réponses aux QCM avec leurs fréquences et le Tableau 6.7 contient les statistiques des réponses sur des échelles de type Likert. Ces échelles sont toutes à 5 niveaux avec des valeurs entières allant de -2 à 2.
En conclusion sur cette évaluation, on peut dire, malgré le faible nombre de réponses, que HeTOP est un outil réellement multi-usages : il est autant utilisé pour la consultation des classifications que pour l'enseignement ou l'indexation. On peut par contre dire que la traduction est aujourd'hui son utilisation principale, notamment pour le MeSH qui est consulté dans 96,9% des cas. Les commentaires libres des utilisateurs confirment cela puisque beaucoup d'entre eux expriment leur intérêt vis-à-vis de l'accès à PubMed, qu'il soit direct via l'onglet dédié ou le Constructeur de Requêtes ou indirect par un simple copié/collé des libellés anglais. On peut également dire que l'utilisation de HeTOP est ponctuelle pour une grande majorité d'utilisateurs (82,9% d'entre eux n'accèdent à HeTOP que moins d'une heure par semaine). Enfin, la qualité du site générale et sur la recherche d'information semble satisfaisante pour les utilisateurs (1,1 et 1 en moyenne sur l'échelle de -2 à 2) mais les éléments affichés et les fonctionnalités restent encore assez complexes (0,3 et 0,5 en moyenne).
Nous espérons effectuer plus d'enquêtes dans un futur proche pour explorer des pistes améliorant ces points et consolidant les fonctionnalités importantes.

5. https://cispro.chu-rouen.fr/cep/#quest=hetop, envoyé par courriel aux utilisateurs inscrits et lien disponible depuis la page d'accueil

Question	Réponse	Pourcentage
Pour quel(s) objectif(s) utilisez-vous HeTOP ?	Traduction	65,6%
	Consultation des classifications	37,5%
	Connaissance (enseignement)	35,9%
	Indexation/codage	32,8%
	Consultation de l'interopérabilité des classifications (alignements)	12,5%
En moyenne, combien de temps utilisez vous HeTOP par semaine ?	Moins de 10 minutes	46,9%
	De 10 à 30 minutes	26,6%
	De 30 minutes à 1 heure	9,4%
	De 1 à 5 heures	10,9%
	Plus de 5 heures	6,2%
Le(s)quel(s) de ces onglets utilisez-vous fréquemment dans HeTOP, pour un concept choisi ?	Description (onglet principal)	87,5%
	Hiérarchies	46,8%
	Ressources	29,7%
	Relations	28,1%
Parmi les classifications ci-dessous, lesquelles utilisez-vous fréquemment ?	MeSH/CISMeF	96,9%
	CIM-10	28,1%
	SNOMED int. 3.5	23,4%
	Autres	34,4%

TABLE 6.6 – Tableau des principales réponses aux QCM d'évaluation de HeTOP

Question	Valeurs réponses min/max (-2 à 2)	Moyenne	Écart-type
La qualité des résultats par rapport à votre recherche dans HeTOP est globalement :	très peu satisfaisante / peu satisfaisante / moyenne / satisfaisante / très satisfaisante	1	0,7
Comment qualifierez-vous la difficulté d'utilisation du site HeTOP (recherche, navigation, etc.) ?	très complexe/../simple	0,5	0,9
Comment qualifierez-vous la difficulté du contenu de HeTOP (vocabulaire, métadonnées, etc.) ?	très complexe/../simple	0,3	0,8
La qualité générale du site est :	très peu/../très satisfaisante	1,1	0,6

TABLE 6.7 – Résultats des questions d'évaluation de HeTOP à échelle de Likert

6.5 Comparaison avec des systèmes proches

Après avoir choisi d'implémenter notre propre système 3M, il a fallu le comparer aux principaux systèmes proches. Outre les différences méthodologiques, fonctionnelles et techniques évidentes, j'ai voulu ici pointer du doigt les différences notables sur certains critères objectifs pouvant montrer les forces et les faiblesses de chaque outil. Cela permet non seulement de situer nos applications et nos méthodes mais aussi d'éventuellement de les améliorer.
Dans ce cadre, nous avons travaillé sur une étude visant à comparer HeTOP et BioPortal, autant sur le fond que sur la forme, en établissant une liste de critères de comparaison entre portails termino-ontologiques. Les critères sont rassemblés en cinq groupes [Grosjean *et al.*, 2014] :

1. le contenu : nombres d'éléments intégrés (SOC, concepts, termes, langues, relations, etc.) ;
2. la politique et la communauté : choix d'intégration et de mises à disposition des contenus et technologies, participation de la communauté, etc. ;
3. les fonctionnalités et les outils : moteur de recherche, types d'affichages, multilinguisme, etc. ;
4. l'interface graphique et l'utilisabilité : interactions avec les utilisateurs et temps de réponses ;
5. les méthodes et technologies : modèles de données, API, ré-utilisabilité, etc.

Je vais donc m'appuyer sur ces critères pour comparer les quatre systèmes présentés dans l'état de l'art à HeTOP et ainsi discuter des avantages et inconvénients de chacun d'eux.

6.5.1 L'UMLS

Contenu

En terme de contenu, l'UMLS est une source d'information extrêmement importante autant par son volume (89 SOC, 2 930 638 CUI, 11 399 740 termes sur 21 langues et 84 306 194 relations) que sa qualité. Chacun des 9 645 410 concepts est rattaché à un CUI.
L'apport principal de l'UMLS en terme d'enrichissement de contenu aux SOC intégrés est uniquement lié au fait de relier chaque terme à un CUI et chaque CUI à un Type Sémantique.
Part ailleurs, l'UMLS propose l'outil MetaMap[6] pour créer des alignements automatiques ou encore repérer des concepts dans des textes. L'entité `MRMAP` stocke d'ailleurs un certain nombre d'alignements automatiques et manuels entre concepts

6. http://metamap.nlm.nih.gov/

terminologiques au sein de l'UMLS. Cependant, leurs origines et leurs qualités sont variables et souvent indéterminées.

Politique et communauté

Il est difficile d'établir clairement les critères nécessaires à l'intégration d'un nouveau SOC dans l'UMLS. Cependant, il est évident que les SOC déjà présents sont fortement utiles pour la communauté bio-médicale et en particulier pour la communauté américaine. En effet, la NLM, qui développe et maintient l'UMLS, dispose de multiples partenariats et donc d'utilisations différentes de l'UMLS ; ses efforts se concentrent donc sur les SOC en anglais et standards aux États-Unis et au Canada : LOINC, MeSH, SNOMED CT, FMA, etc. Quand ces SOC existent dans d'autres langues, ils sont également disponibles.
En terme de mise à disposition, l'UMLS est particulièrement ouverte puisqu'il est possible de télécharger l'intégralité du méta-thésaurus. Cependant, son utilisation est cadrée par l'adhésion à une licence globale mais également à un certain nombre de permissions spécifiques à chaque SOC.

Modèle

L'UMLS présente le plus grand nombre de caractéristiques communes au S3M. Par définition, l'UMLS est multi-terminologique mais également multi-lingue (et inter-lingue). Bien qu'il se définisse comme un méta-thésaurus de terminologies bio-médicales, une analyse du modèle de données permet de penser qu'il serait également multi-discipline. En effet, l'UMLS propose peu d'entités : elle sépare la partie dédiée à la déclaration des concepts (`MRCONSO`) de leurs définitions (`MRDEF`) et de leurs relations (`MRREL`, `MRHIER` et `MRMAP`). Cela est fortement comparable à notre méta-modèle terminologique 3M. Il n'y a aucune raison pour que ce modèle ne soit pas également le support de SOC hors Santé (aucune entité ou propriété n'est dédiée à une discipline particulière). Par ailleurs, le modèle de l'UMLS propose une partie spécifique au méta-thésaurus avec notamment les entités `MRCUI` et `MRSTY` qui définissent les concepts uniques et les types sémantiques. Il s'agit là d'une particularité de l'UMLS qui centre vraiment l'information autour des concepts UMLS et donc, d'une vue plus large, plus *unifiée* des SOC. C'est donc là le point le plus divergent avec notre vision multi-terminologique. Notre approche ne consiste pas à créer un méta-SOC mais seulement de les stocker de façon parallèle en créant des ponts via l'interopérabilité. Même si l'UMLS introduit des traitements TAL dans leurs alignements, le principal travail des experts consiste à rattacher (à la main) un CUI à chaque concept de chaque SOC pour le faire entrer dans ce méta-thésaurus. C'est une opération extrêmement couteuse, impossible à mettre en place avec des moyens limités. De plus, créer une telle structure n'est pas sans dangers. C'est d'ailleurs l'objet du travail de Merill [Merrill, 2009], qui expose différents problèmes classiques

concernant la « synonymie relative » ou encore la polysémie. Bodenreider montre également différents problèmes de cycles générés par l'agrégation de termes de niveaux hiérarchiques différents ou de termes composés dans un même concept UMLS [Bodenreider, 2001]. La gestion de ces concepts (au sens CUI) devient alors de plus en plus complexe à mesure que le méta-thésaurus s'étoffe non seulement parce que cette étape reste subjective mais aussi parce que la distinction des sens des concepts devient floue avec l'ajout de « faux » synonymes. Certaines erreurs apparaissent donc et le phénomène de cascades d'erreurs (effet « boule de neige » décrit dans le S3M : cf. 6.2.2) est également vrai pour l'UMLS.
La base de données de l'UMLS est relationnelle et ne permet donc pas de stocker des ontologies formelles de façon native. La méthodologie pour intégrer le lexique d'ontologies au sein de l'UMLS est la même que celle employée dans le S3M ; les ontologies sont « dégradées » pour entrer dans un modèle terminologique. Cependant, plusieurs initiatives ont vu le jour récemment pour palier ce problème et proposent des schémas de BDD relationnelles permettant le stockage d'ontologies formelles [AlAmri, 2012].
Enfin, comme dans le S3M, chaque version « majeure » de SOC constitue un SOC différent et les versions sont donc stockées conjointement dans le système.

Portail et outils

L'UMLS propose un outil en ligne pour consulter son méta-thésaurus. Ce portail, appelé UTS [7], permet donc de rechercher des concepts et d'afficher le contenu de l'UMLS via une interface web. Il permet également de télécharger diverses sources de SOC et d'utiliser un Service Web.
En premier lieu, la recherche ne peut se faire qu'en anglais, seulement sur les PT et seulement en simple troncature (contre double troncature dans HeTOP). La recherche est rapide mais l'affichage des informations des concepts est relativement lente (plus de 3 secondes en moyenne). Ces informations sont en fait agrégées pour chaque concept UMLS : tous les éléments des différents SOC sont rassemblés tels qu'ils apparaissent dans la base de données. Les méta-données ne sont donc pas explicitées (libellés souvent inintelligibles) et sont affichés pèle-mêle.
En outre, il est possible de restreindre les recherches à certains SOC en particulier mais les résultats des recherches ne sont pas séparés par SOC. Le navigateur de l'UTS est en fait un outil exploitant le méta-thésaurus et son réseau sémantique riche. En cela, il est donc différent de HeTOP puisque ce dernier offre une vision multi-terminologique et non pas une vision unifiée. De la même façon, l'UTS n'a pas pour vocation d'exploiter les différents usages natifs des SOC, comme la recherche documentaire pour le MeSH, par exemple.

7. Pour *UMLS Terminology Services* : https://uts.nlm.nih.gov/

6.5.2 BioPortal

Contenu

BioPortal contient plus de 360 SOC, ce qui correspond à 5 960 457 concepts et 6 600 000 termes. Malheureusement, il n'a pas été possible de recueillir le nombre relations excepté le nombre de 5 000 000 d'alignements exacts générés ; il s'agit d'ailleurs du seul apport de BioPortal en terme d'ajout de contenu.

Politique et communauté

BioPortal est conçu comme un dépôt d'ontologies bio-médicales. À ce titre, il propose aux utilisateurs de téléverser eux-mêmes leurs SOC dans un format standard (OWL ou OBO). Il est alors possible de déterminer si ces SOC sont ouverts au public ou seuls certains peuvent y accéder. Cependant, beaucoup de SOC intégrés proviennent de l'UMLS ou de la OBO Foundry [8].
BioPortal propose des API ouvertes, des Services Web, une documentation riche et même la possibilité d'installer sa propre instance de l'outil.

Modèle

BioPortal ne repose pas sur un méta-modèle mais sur une collection d'ontologies stockées dans un entrepôt RDF. Du fait de leur politique qui consiste à n'intégrer que des SOC aux formats ontologiques, les imports sont largement facilités. Quelques ajustements peuvent cependant avoir lieu si certains éléments ne sont pas bien représentés.
En terme de versionnage, BioPortal propose une gestion très fine des versions des SOC avec la possibilité de consulter chaque concept dans chaque version disponible. Enfin, BioPortal ne gère pas le multi-linguisme mais propose des « vues » sur des versions de SOC dans d'autres langues. Il existe encore quelques anomalies à ce sujet car le NCBO travaille actuellement sur une intégration d'autres langues que l'anglais.

Portail et outils

BioPortal est avant tout l'outil en ligne que connaissent les utilisateurs. Bien qu'il soit spécialement conçu pour consulter les ontologies qu'il contient, le portail souffre de lenteurs importantes : notre étude a montré que les temps de réponse étaient bien supérieurs à ceux de HeTOP (pour un nombre de termes sensiblement le même) et que le temps d'affichage d'un concept de SOC était en moyenne de 4,1 secondes. Plusieurs témoignages d'utilisateurs ont démontré qu'il était très difficile de se servir quotidiennement de BioPortal dans des contextes de codage ou d'indexation.

8. Banque de donnés d'ontologies bio-médicales : http://www.obofoundry.org/

En outre, BioPortal offre toutes les fonctionnalités de base de consultation et de recherche dans les SOC. Cependant, il ne gère pas la poly-hiérarchie ou encore l'inter- et multi-linguisme. Enfin, l'accès aux ressources ne peut pas se faire directement dans l'outil mais il faut passer par une application dédiée.

6.5.3 EBI OLS

Contenu

OLS contient actuellement 93 ontologies soit 2 223 753 de termes. Aucun enrichissement n'est appliqué dans OLS. Les SOC sont insérés de façon indépendante les uns aux autres, sans aucune interopérabilité.

Politique et communauté

Le OLS est avant tout un outil pour les biologistes et bioinformaticiens. À ce titre, son utilisation relève essentiellement de recherches et d'extractions de codes via des Services Web. OLS propose donc des API et des SW pour fournir ce type d'information mais également un outil en ligne pour consulter les termes des SOC intégrés.
Le choix de l'intégration des SOC est donc uniquement lié à leurs utilisations dans des pipelines d'analyses classiques des différents domaines de la bioinformatique (génomique, protéomique, etc.). Ces SOC sont extrêmement précis (à une espèce comme la *Drosophilia Phenotype Ontology*, à une étude comme la *Spider Comparative Biology Ontology*) et souvent peu volumineux, à part quelques cas (moyenne de 23 911 termes par SOC mais avec un écart-type de 129 353 termes).
L'une des grandes forces de cet outil est sa fréquence de mise à jour quotidienne qui assure aux utilisateurs une fiabilité de l'information.

Modèle

Le modèle de l'OLS est basé sur le schéma de base de données BioSQL. Il s'agit donc d'une base de données relationnelle alimentée par un programme générique parcourant des fichiers OBO. Il existe en fait un méta-modèle de vocabulaire contrôlé définissant des `Termes` impliqués dans des relations (dont des hiérarchies) et pouvant avoir des attributs. Ce modèle n'est pas multi-lingue. En outre, les noms de propriétés sont fortement liés aux utilisations bioinformatiques des SOC (associations avec des séquences géniques, etc.).
Il n'est pas fait mention de versionnage dans ce modèle. La stratégie adoptée est celle de l'écrasement à chaque nouvelle mise à jour de SOC.

Portail et outils

L'interface web de l'OLS propose de ne sélectionner qu'un seul SOC à la fois pour effectuer des recherches. Il est cependant théoriquement possible de le faire via un Service Web. La recherche s'effectue sur les PT ou les identifiants mais ne peut se faire que via un mécanisme d'auto-complétion avec une troncature à droite seulement. De plus, on ne peut avoir qu'un seul résultat à chaque recherche. Comme expliqué avant, OLS ne traite que les données en anglais.
Pour un concept de SOC donné, l'affichage se fait en un tableau unique avec toutes propriétés et leurs valeurs listées sans aucune explicitation. L'outil n'est pas conçu pour faire de la recherche documentaire mais propose tout de même des identifiants d'équivalence dans d'autres SOC ou bases de données (non cliquables cependant).

6.5.4 LexGrid et le NCI Term Browser

Contenu

Le NCI Term Browser (NTB) contient actuellement 22 SOC axés essentiellement autour de ressources d'annotation pour les DPI (LOINC, CIM-9, MedDRA, SNOMED CT, ...) et pour la recherche fondamentale en biologie moléculaire (Gene Ontology, ChEBI, ...). Le SOC le plus mis en avant (sélection par défaut) est le NCIT[9] qui est le thésaurus développé par la communauté du Centre National du Cancer aux États-Unis. Il s'agit d'un référentiel en oncologie (et domaines transversaux) dont l'utilisation dépasse aujourd'hui les frontières américaines [de Coronado *et al.*, 2009].
Les différentes statistiques sur la volumétrie ne sont pas accessibles pour LexEVS mais on peut estimer que le nombre de concepts terminologiques dépasse les 2 000 000.

Politique et communauté

LexGrid a été créé pour soutenir un standard, aussi bien au niveau du modèle terminologique qu'au niveau des outils de communications (API et SW). En cela, LexGrid est ouvert et dispose d'une documentation fournie.
Le portail NTB n'est qu'une vitrine du serveur LexEVS mais propose tout de même des fonctionnalités de base et se veut un démonstrateur à part entière du système. Nous l'avons donc évalué en tant que tel. Au niveau des droits sur les SOC, le NTB propose d'adhérer à une charte (*disclaimer*[10]) globale puis à une charte par SOC à chaque connexion.

9. *National Cancer Institute Thesaurus* : http://ncit.nci.nih.gov/
10. « Avis de non responsabilité »

Modèle

LexGrid est en fait un ensemble de sous-modèles. En effet, il est composé de 10 sous-schémas définissant chacun un domaine (`concepts`, `codingSchemes`, `relations`, `services`, etc.) [11]. Si l'on veut appréhender LexGrid en entier, il faut gérer pas moins de 79 entités différentes reliées entre elles. Ces entités permettent la gestion fine de certains éléments (versions, types de données, langues, types de relations, etc.). Tout ceci a pour but de standardiser non seulement le vocabulaire employé pour définir des SOC mais aussi pour les stocker et faire fonctionner les programmes les exploitant. Son implémentation est en donc peu facilitée.
Ce modèle est multi-terminologique, multi-lingue et multi-domaine. Il calque exactement avec les besoins du S3M. Cependant, sa complexité et son manque de pragmatisme en font un outil peu flexible et difficilement implémentable dans un environnement fonctionnel. La documentation est souvent peu précise et demande de multiples renvois. De plus, il n'existe pas de schéma de base de données directement importable.

Portail et outils

Le NTB propose des fonctionnalités de recherche avec ou sans troncatures sur les PT, identifiants et sur les propriétés. La recherche est multi-terminologique et rapide. Cinq onglets composent la description des concepts : (i) le détail sur les attributs, (ii) la liste des synonymes, (iii) les relations, (iv) les alignements et (v) une vue agrégeant l'intégralité des 4 premiers onglets. Il est possible de naviguer entre concepts de SOC différents mais les libellés de relations sont souvent inintelligibles. Une autre fenêtre permet de visualiser les hiérarchies dans un arbre interactif. Le tout est fluide et adapté à la consultation rapide.
Bien que LexGrid soit multi-lingue, le NTB ne propose pas de contenu dans d'autres langues que l'anglais.

6.5.5 Synthèse des comparaisons

Comme nous venons de le voir, le S3M et HeTOP ont tous deux des équivalents de poids dans le domaine, autant sur le fond que sur la forme.
Concernant le modèle 3M, l'UMLS et LexGrid proposent deux approches assez divergentes de la notion de méta-modèle terminologique et de son implémentation. En effet, alors que l'UMLS dispose d'un méta-modèle compact et orienté méta-thésaurus, LexGrid se positionne comme un standard terminologique via un modèle formel complet mais éclaté. Les deux modèles sont 3M mais supposent des implémentations très différentes et donc, des investissements en temps différents. Côté

11. Voir http://informatics.mayo.edu/LexGrid/downloads/LexGrid%20Model/schemas/2005/01/EAwebpublish/index.htm

EBI OLS, le modèle issu de BioSQL est très compact et très proche du S3M mais trop spécifique à la biologie et ne gère pas le multi-linguisme. Le modèle le plus éloigné de ceux-là est donc celui de BioPortal qui, en fait, n'en définit pas formellement ; il s'agit d'un entrepôt RDF et comme tous les SOC intégrés dans BioPortal possèdent des formats ontologiques standards, cela ne pose aucun problème.
Le S3M est évidemment un modèle de plus (même si très proche des méta-modèles classiques) mais possède une simplicité et une généricité importante et pragmatique. Son implémentation est simple et intuitive. L'approche LexGrid nécessite trop de temps pour appréhender le système et des fonctionnalités beaucoup trop contraignantes. La maintenance et le développement des couches métier sont des points noirs pour l'utilisation de ce type de modèle. De son côté, BioPortal possède une approche inverse en terme de modélisation. Il s'agit d'un choix stratégique discutable vis-à-vis des performances, mais cela est sans doute la rançon de l'utilisation d'une technologie encore jeune. Je reste d'ailleurs convaincu que les entrepôts RDF sont de sérieux candidats à la base de données de demain. Stocker tous les SOC sous forme de triplets est sans doute la meilleure solution à terme même si les outils d'aujourd'hui ne les exploitent pas encore au mieux. Néanmoins, notre choix s'est porté sur un méta-modèle 3M compact et implémentable dans une base de données relationnelle. Il s'agissait avant tout d'une approche pragmatique, rapide et facile à mettre en place, à développer et à maintenir. L'objectif principal du S3M est de répondre très rapidement aux utilisateurs et d'être également assez souple pour intégrer n'importe quel SOC dans n'importe quelle langue.
Lorsque l'on compare le contenu des serveurs 3M présentés ici, il faut séparer les systèmes qui ne peuvent être alimentés que par des sources standards (BioPortal, LexGrid et EBI OLS), des systèmes qui s'adaptent et nécessitent une intervention humaine, autant sur l'informatique que sur l'expertise du contenu (HeTOP et UMLS). Dans ce contexte, les coûts de gestion sont bien différents : un système autonome nécessitera moins de temps à la maintenance et à la flexibilité. Dans notre cas, HeTOP est avant tout une somme de projets, d'interventions humaines et de technologies. L'UMLS est également une source d'une grande valeur mais reste attaché à son contenu plus qu'à son contenant et ne s'intéresse que très peu à l'utilisation première et native de chaque SOC. De plus, le coût humain à l'élaboration et à la maintenance d'un tel méta-thésaurus est énorme. C'est pour cela que notre choix s'est porté sur un serveur multi-terminologique et non un méta-thésaurus. Cela dit, le S3M offre de multiples enrichissements de contenu, souvent spécifiques et parfois issus de projets de recherche novateurs (icônes VCM, etc.). Cependant, ces ajouts de contenu constituent une valeur-ajoutée importante qu'il est parfois difficile de diffuser. Le S3M se distingue d'ailleurs des autres systèmes par sa politique, moins ouverte et ne permettant donc pas d'exporter facilement les données.
Pour conclure sur le modèle 3M, le S3M présenté dans ces travaux est donc très

proche des modèles connus et classiques de multi-terminologie. Il se veut compact et flexible et simple à implémenter. Il est également à noter qu'une norme est sortie fin 2011 sur les « thésaurus pour la recherche documentaire » (ISO 25964). Bien que la documentation soit payante, des informations confirment que les thésaurus sont définis par un ensemble de concepts et de termes, éventuellement regroupés par des groupes de concepts (cf. Figure Annexe A.1). Notre modèle est donc proche et même compatible avec cette représentation.

Concernant les portails, on peut remarquer que certaines fonctionnalités de base de consultation et de recherche des SOC font consensus. Cependant, que ce soit par les performances ou par la présentation, de nombreux aspects font diverger les outils présentés ici.
HeTOP a été créé dans une optique bien précise, selon une approche *bottom-up*, classique mais pragmatique. Ainsi, HeTOP se veut lui-même un outil prêt à l'emploi, par son Service Web mais surtout par son site web dédié à l'humain. Plus précisément, HeTOP est dédié à l'utilisateur, presque exclusivement à l'expert ; certaines méta-données, même explicitées n'ont que peu d'utilité pour la majorité des utilisateurs mais les chercheurs ou les experts de ces SOC peuvent y voir un intérêt tout particulier. HeTOP est donc le seul portail a offrir des méta-données explicitées et multi-lingues. Les autres portails décrits affichent tous les éléments des SOC de façon brute et native sans aucun enrichissement (à part quelques alignements automatiques). De plus, l'UTS est peu adaptée dans des utilisations quotidiennes car s'attache avant tout à décrire les concepts du méta-thésaurus et non pas à ceux des SOC en particulier. Pour BioPortal, il s'agit en plus d'un problème récurrent de performances et la multi-terminologie est peu exploitée (et même pas du tout pour OLS par ailleurs).
Côté performances, HeTOP et NTB offrent de bons résultats, acceptables pour les utilisateurs au quotidien. J'exclus de cette comparaison UTS et OLS qui n'offrent pas vraiment de vision multi-terminologique et donc une recherche adaptée sur plusieurs SOC à la fois. Par ailleurs, HeTOP est le seul système à gérer la recherche multi-lingue.
En conclusion sur la comparaison des portails, nous pouvons affirmer que le développement de HeTOP était essentiel car aucun outil existant à notre connaissance n'offre ses fonctionnalités, axées sur la compréhension par l'humain et assurant une utilisabilité correcte pour une pratique quotidienne. Sa gestion du multi- (et inter-) linguisme, son contenu à valeur ajoutée et ses fonctionnalités documentaires en font un outil à part entière pour la communauté.

Chapitre 7

Conclusions et perspectives

Dans le cadre de ce travail, nous nous sommes intéressés à l'élaboration d'un modèle multi-terminologique, multi-discipline, multi-lingue pouvant supporter l'intégration de nombreux SOC, quelque soient leurs tailles et leurs modèles. Ce méta-modèle devait donc être robuste et flexible à la fois afin d'y accueillir des données hétérogènes reliées entre elles. Son implémentation devait être rapide et modulaire en créant un serveur référent contenant ces SOC et agissant comme une véritable plateforme utile à la communauté. De ce fait, l'outil central de cette étude était la réalisation d'un site web et d'un Service Web exploitant le S3M et pouvant être à la fois utiles pour les humains et les machines. Ce portail devait être adapté à une utilisation quotidienne, à grande échelle et personnalisable. Les objectifs de cette application étaient donc d'aider à l'indexation, au codage, à l'enseignement et à la gestion terminologique.

Pour réaliser cela, il a donc fallu étudier des systèmes existants proches et partagés par les communautés de l'informatique médicale et de l'ingénierie des connaissances. Après études et comparaisons approfondies, nous avons choisi de créer notre propre méta-modèle 3M suite, notamment, au projet InterSTIS. Son implémentation dans un modèle logique de données générique a été un défi important, autant sur le niveau méthodologique que technique. Ce modèle de base de données a montré une puissance plus grande que nous l'avions imaginée et a donc été utilisé à grande échelle dans le SI de CISMeF et dans d'autres projets depuis.

La suite et fin des travaux ont été la conception et le développement du portail et du Service Web exploitant le S3M. HeTOP est maintenant en production depuis plus d'un an et accuse un trafic journalier régulier et, au vue des enquêtes, son utilisation est très prisée par de nombreux spécialistes et étudiants. De plus, cet outil constitue un atout important pour les activités de recherche de CISMeF mais également de tous ses partenaires ; en effet, HeTOP affiche et met en valeur les différents travaux autour des SOC, que ce soit autour de l'interopérabilité, de la RI, de l'enseignement ou de la connaissance médicale en général. Enfin, les différentes déclinaisons de HeTOP et les outils d'édition de SOC constituent aujourd'hui des applications

de production utilisée quotidiennement, notamment au CHU de Rouen (HeTOP SFMU pour le codage des DPI ou encore les flux de terminologies dans le circuit de prescription de la biologie à l'hôpital).

Cependant, plusieurs limites se posent sur notre modèle 3M et sur l'utilisation de nos outils.
Premièrement, le versionnage actuel n'est pas entièrement satisfaisant car il est souvent compliqué voire impossible de retracer finement les changements de versions des informations terminologiques. Par ailleurs, à chaque mise à jour de SOC, il est très difficile d'évaluer les impacts possibles sur les changements de libellés ou de hiérarchies des concepts : alignements, synonymes, etc.
En outre, quelques cas de représentation d'éléments posent encore problème dans notre modèle : créer des relations *n-n*, ajouter des propriétés sur des propriétés ou encore effectuer de la post-coordination de codes à plus de 2 éléments.
De plus, pour des raisons essentiellement politiques, HeTOP et le S3M ne permettent pas les extractions directes des SOC, de leurs alignements ou de leurs traductions. D'ailleurs, la question du(des) format(s) d'export se poserait étant données les limites actuelles du SKOS et des difficultés de transformation en OWL des terminologies.
En ce qui concerne HeTOP, les évaluations par les utilisateurs ont permis de lever quelques limites comme des lourdeurs dans l'ergonomie et la complexité de certains libellés et contenus. Du travail reste en outre à effectuer pour assurer une bonne utilisation de chaque SOC de façon spécifique. Un gros travail est déjà été fait pour le MeSH étant donné qu'il s'agit de la terminologie pivot de l'équipe mais d'autres SOC mériteraient plus d'attention. Enfin, HeTOP est encore trop centré sur le français alors qu'il se veut multi-lingue. Certaines langues comme le japonais ou l'arabe ont des mécanismes linguistiques totalement différents des langues latines et leur traitement informatique n'est aujourd'hui pas adapté dans HeTOP.

J'ai pu étudier beaucoup de SOC (plus de 50) très variés, autant en structures, qu'en modèles, qu'en volumétrie, utilisation, origine, date de création, etc. Que ce soit une simple liste de termes ou une ontologie, la modélisation d'un SOC reste un point central dans sa conception et son utilisation. Je ne pense pas qu'il existe une seule « bonne façon » de modéliser un SOC car cela dépend, selon moi, essentiellement de son utilisation. Ces travaux de thèse m'ont, je pense, beaucoup éclairé quant à la façon la plus adéquate de décrire un modèle de SOC, et ce, quelque soit son niveau de complexité. À l'ère de l'ontologie, dans le domaine de l'ingénierie des connaissances, énormément de nouveaux concepteurs de SOC sont convaincus que d'utiliser les ontologies et la meilleure manière de créer un nouveau SOC. Comme le démontrent certains éléments de ces travaux de thèse, je ne pense pas que l'ontologie soit au-

jourd'hui le type de SOC le plus adapté. Gagner en complexité et en formalisme ne signifie pas gagner en performance et en utilisabilité, bien au contraire. Les ontologies sont des outils extrêmement puissants permettant de structurer la connaissance, de raisonner et d'inférer des faits, de représenter des règles, etc. Cependant, les algorithmes s'appuyant sur les ontologies souffrent encore, pour beaucoup, de faibles performances (temps de calculs de raisonnement sur des ontologies formelles de plus de 80 000 classes comme la FMA, par exemple). De plus, la complexité et le formalisme des ontologies nécessitent une expertise poussée afin d'en tirer le meilleur parti. C'est une des raisons pour lesquelles le choix du niveau terminologique s'est imposé lors de la conception du modèle 3M.
Côté technique, beaucoup de technologies, langages et méthodologies ont été abordées ; le travail sur les fichiers, leurs formats et leurs complexités, les bases de données, les outils de modélisation, les ontologies, les Services Web, les RIA, etc. Cette étude offre un panel de problématiques recoupant divers thèmes de l'informatique et de la recherche en général.

Malgré le travail effectué, beaucoup d'éléments restent à approfondir, surtout du point de vue opérationnel. Le S3M est un cadre de travail et beaucoup de nouveaux SOC intéressants pourront y être intégrés prochainement. Des travaux autour de ces SOC sont aussi en cours et pourront bientôt voir leurs résultats dans HeTOP.
Par ailleurs, le S3M reste, selon nous, encore trop centré sur lui-même. Nous espérons prochainement le rendre plus visible, non seulement via HeTOP mais aussi via son SW. En outre, il existe aujourd'hui des standards de communication en terminologie (CTS2), complexes mais puissants pouvant permettre à des systèmes de communiquer facilement. Nous aimerions travailler sur ce point.
En plus de cela, la question du versionnage reste central ; gérer les différentes versions d'un même SOC le plus finement possible est un défi important. Nous souhaitons également automatiser certaines mises à jour de SOC via la mise en place de flux de fichiers sources officiels (quelques SOC le permettent comme la CCAM, la LPP, etc.).
En outre, il nous reste encore beaucoup de travail autour de HeTOP, autant sur son ergonomie que sur ses fonctionnalités. Nous voulons notamment améliorer le « Constructeur de requête », de plus en plus prisé par les utilisateurs. Nous avons d'ailleurs déjà lancé une campagne d'évaluation sur son utilisabilité et sa pertinence. Par ailleurs, des efforts sont à produire sur la rédaction de documentations, autant techniques que fonctionnelles.
En ce qui concerne la partie édition de SOC, nous souhaitons également améliorer nos outils, qu'il s'agisse des validations et des saisies manuelles : du travail reste à effectuer concernant la mise à jour en temps réel des hiérarchies de SOC par exemple. En terme de contenu des SOC, afin de limiter les impacts des synonymes peu perti-

nents, nous avons récemment entrepris d'élaborer une échelle (ou catégorisation) des termes en fonction de leur qualité, de leur type et de leur représentativité du concept auquel ils sont rattachés. Les différentes applications seront alors configurées pour n'utiliser que telle ou telle catégorie, en fonction de leur niveau de criticité (niveau précision/bruit attendu et acceptable) ; il s'agit d'un projet ambitieux mais pouvant avoir un impact très important sur l'indexation automatique et la RI en général.
Enfin, nous avons également pour objectif d'améliorer notre modèle logique générique de données. Ces résultats sont très prometteurs mais quelques pistes restent à explorer : l'historisation notamment mais aussi la création de groupes de relations.
Nous souhaitons également continuer à le diffuser plus largement, du point de vue scientifique d'abord avec un article en cours d'écriture, et du point de vue opérationnel avec des implémentations pour de nouveaux projets.
Les différentes améliorations du S3M et de HeTOP seront apportées via certains projets de recherche comme SYNODOS (cf. 5.4.3 et 5.5.4) pour la création d'un SOC spécifique et l'indexation automatique mais aussi grâce au projet ADR-Prism [1] pour l'exploitation des SOC.
Le principal projet qui devrait nous permettre de réaliser une partie importante de ces perspectives est le projet PlaIR II déjà financé par le conseil régional de Haute-Normandie.

1. Financé par le Fonds unique interministériel, ADR-Prism (pour *Adverse Drug Reactions from Patient Reports In Social Medias*) a pour objectif de créer une source de connaissances encore inexploitée : les messages des patients dans les forums concernant les effets indésirables des médicaments

Bibliographie

AlAmri, A. : The relational database layout to store ontology knowledge base, in *2012 International Conference on Information Retrieval Knowledge Management (CAMP)*, 2012, (pp. 74–81).

Balani, N., & Hathi, R. : *Apache CXF Web Service Development : Develop and Deploy SOAP and RESTful Web Services*, Packt Publishing Ltd, 2009.

Binding, C., & Tudhope, D. : KOS at your service : Programmatic access to knowledge organisation systems, *Journal of Digital Information*, 4(4).

Bizer, C., Heath, T., & Berners-Lee, T. : Linked data - the story so far :, *International Journal on Semantic Web and Information Systems*, 5(3), (2009), 1–22.

Bodenreider, O. : Circular hierarchical relationships in the UMLS : etiology, diagnosis, treatment, complications and prevention., *Proceedings of the AMIA Symposium*, (pp. 57–61).

Bodenreider, O. : The unified medical language system (UMLS) : integrating biomedical terminology, *Nucleic acids research*, 32(Database issue), (2004), D267–270.

Carpentier, M., Saliou, P., Le Du, I., Le Guillou, C., Lecornu, L., & Cauvin, J. M. : Détection et évaluation des associations entre les nomenclatures CCAM et LPP, brest, *Revue d'Épidémiologie et de Santé Publique*, 58, Supplement 1, (2010), S5.

Charlet, J., Bachimont, B., & Jaulent, M.-C. : Building medical ontologies by terminology extraction from texts : An experiment for the intensive care units, *Computers in Biology and Medicine*, 36(7–8), (2006), 857–870.

Cimino, J. J., & Zhu, X. : The practical impact of ontologies on biomedical informatics, *Yearbook of medical informatics*, (pp. 124–135).

Côté, R., Reisinger, F., Martens, L., Barsnes, H., Vizcaino, J. A., & Hermjakob, H. : The ontology lookup service : bigger and better, *Nucleic acids research*, 38(Web Server issue), (2010), W155–160.

Côté, R. G., Jones, P., Apweiler, R., & Hermjakob, H. : The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries, *BMC bioinformatics*, 7, (2006), 97.

Darmoni, S. J., Pereira, S., Névéol, A., Massari, P., Dahamna, B., Letord, C., Kedelhué, G., Piot, J., Derville, A., & Thirion, B. : French infobutton : an academic and... business perspective, in *AMIA Symp.*, IOS Press, 2008, (p. 920).

Darmoni, S. J., Soualmia, L. F., Letord, C., Jaulent, M.-C., Griffon, N., Thirion, B., & Névéol, A. : Improving information retrieval using medical subject headings concepts : a test case on rare and chronic diseases, *Journal of the Medical Library Association : JMLA*, 100(3), (2012), 176–183.

de Coronado, S., Wright, L. W., Fragoso, G., Haber, M. W., Hahn-Dantona, E. A., Hartel, F. W., Quan, S. L., Safran, T., Thomas, N., & Whiteman, L. : The NCI thesaurus quality assurance life cycle, *Journal of Biomedical Informatics*, 42(3), (2009), 530–539.

Degoulet, P., & Fieschi, M. : *Informatique médicale*, Elsevier Masson, 1998.

Dinu, V., & Nadkarni, P. : Guidelines for the effective use of entity–attribute–value modeling for biomedical databases, *International Journal of Medical Informatics*, 76(11–12), (2007), 769–779.

Duclos, C., Burgun, A., Lamy, J.-B., Landais, P., Rodrigues, J.-M., Soualmia, L. F., & Zweigenbaum, P. : Le vocabulaire médical, les ressources terminologiques, le codage de l'information en santé, in *Informatique médicale, e-Santé*, (pp. 11–41), Springer Paris, 2013.

Ducrot, H., & Dusserre, L. : *L'Informatique médicale*, Presses Universitaires de France, 1990.

Feigenbaum, E. A., & McCorduck, P. : *The fifth generation : artificial intelligence and Japan's computer challenge to the world*, Reading, Mass. : Addison-Wesley, 1983.

Fung, K. W., & Bodenreider, O. : Utilizing the UMLS for semantic mapping between terminologies, *AMIA Annual Symposium Proceedings*, 2005, (2005), 266–270.

Gaudin, F. : Terminologie : l'ombre du concept, *Meta : Journal des traducteurs*, 41(4), (1996), 604.

Gicquel, Q., Kergoulay, I., Gerbier-Colomban, S., Chariout, S., Bittar, A., Segond, F., Darmoni, S. J., & Metzger, M. : Annotation methods to develop and evaluate an expert system based on natural language processing in electronic medical records, in *MIE*, 2014, accepted.

Girard, A. : *Synonymes françois, leurs différentes significations, le choix qu'il en faut faire pour parler avec justesse*, Wetstein, 1769.

Golbreich, C., Grosjean, J., & Darmoni, S. J. : The FMA in OWL 2, in *AIME*, vol. 6747, Springer-Verlag, 2011, (pp. 204–214).

Golbreich, C., Grosjean, J., & Darmoni, S. J. : The foundational model of anatomy in OWL 2 and its use., *Artif Intell Med*, 57(2), (2013), 119–132.

Grabar, N., Hamon, T., & Bodenreider, O. : Ontologies and terminologies : Continuum or dichotomy ?, *Applied Ontology*, 7(4), (2012), 375–386.

Griffon, N. : *Modélisation, création et évaluation de ux de terminologies et de terminologies d'interface : application à la production d'examens complémentaires de biologie et d'imagerie médicale.*, Ph.D. thesis, Université de Rouen, 2013.

Griffon, N., Kerdelhué, G., Hamek, S., Hassler, S., Boog, C., Lamy, J.-B., Duclos, C., Venot, A., & Darmoni, S. J. : Design and usability study of an iconic user interface to ease information retrieval of medical guidelines, *Journal of the American Medical Informatics Association : JAMIA*.

Griffon, N., Kerdelhué, G., Soualmia, L. F., Merabti, T., Grosjean, J., Lamy, J.-B., Venot, A., Duclos, C., & Darmoni, S. J. : Evaluating alignment quality between iconic language and reference terminologies using similarity metrics, *BMC medical informatics and decision making*, 14, (2014b), 17.

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

Grosjean, J., Merabti, T., Soualmia, L. F., Letord, C., Charlet, J., Robinson, P. N., & Darmoni, S. J. : Integrating the human phenotype ontology into HeTOP terminology-ontology server, *Studies in health technology and informatics*, 192, (2013), 961.

Grosjean, J., Soualmia, L. F., Bouarech, K., Jonquet, C., & Darmoni, S. J. : Comparing BioPortal and HeTOP : towards a unique biomedical ontology portal ?, in *IWBBIO 2014*, 2014, accepted.

Gruber, T. R. : A translation approach to portable ontology specifications, *Knowledge Acquisition*, 5(2), (1993), 199–220.

Guarino, N., & Giaretta, P. : Ontologies and knowledge bases : Towards a terminological clarification, *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing*, (pp. 25–32).

Hodge, G. M. : *Systems of Knowledge Organization for Digital Libraries : Beyond Traditional Authority Files*, Digital Library Federation, Council on Library and Information Resources, 2000.

Humphreys, B. L., Lindberg, D. A. B., Schoolman, H. M., & Barnett, G. O. : The unified medical language system, *Journal of the American Medical Informatics Association : JAMIA*, 5(1), (1998), 1–11.

Joubert, M., Vandenbussche, P.-Y., Dahamna, B., Abdoune, H., Merabti, T., Pereira, S., Boyer, C., Staccini, P., Forget, J.-F., Delahousse, J., Darmoni, S. J., & Fieschi, M. : InterSTIS : Interopérabilité sémantique de terminologies de santé francophones, in P. P. M. Staccini, D. A. Harmel, P. S. J. Darmoni, & P. R. Gouider (Eds.), *Systèmes d'information pour l'amélioration de la qualité en santé*, no. 1 in Informatique et Santé, (pp. 73–83), Springer Paris, 2012.

Kister, L., Jacquey, E., & Gaiffe, B. : Du thesaurus à l'onto-terminologie : relations sémantiques vs relations ontologiques, *CORELA*, 9(1).

Klein, M., Fensel, D., Kiryakov, A., & Ognyanov, D. : Ontology versioning and change detection on the web, in A. Gómez-Pérez, & V. R. Benjamins (Eds.), *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, no. 2473 in Lecture Notes in Computer Science, (pp. 197–212), Springer Berlin Heidelberg, 2002.

Lamy, J.-B. : *Conception et évaluation de méthodes de visualisation des connaissances médicales : Mise au point d'un langage graphique et application aux connaissances sur le médicament*, Ph.D. thesis, Paris 6, 2006.

Lamy, J.-B., Duclos, C., Bar-Hen, A., Ouvrard, P., & Venot, A. : An iconic language for the graphical representation of medical concepts, *BMC Medical Informatics and Decision Making*, 8(1), (2008), 16.

Lamy, J.-B., Duclos, C., Hamek, S., Beuscart-Zéphir, M.-C., Kerdelhué, G., Darmoni, S. J., Favre, M., Falcoff, H., Simon, C., Pereira, S., Serrot, E., Mitouard, T., Hardouin, E., Kergosien, Y., & Venot, A. : Towards iconic language for patient records, drug monographs, guidelines and medical search engines, *Studies in health technology and informatics*, 160(Pt 1), (2010), 156–160.

Lamy, J.-B., Soualmia, L. F., Kerdelhué, G., Venot, A., & Duclos, C. : Validating the semantics of a medical iconic language using ontological reasoning, *Journal of Biomedical Informatics*, 46(1), (2013), 56–67.

Leavitt, N. : Will NoSQL databases live up to their promise ?, *Computer*, 43(2), (2010), 12–14.

Lee, D., Cornet, R., & Lau, F. : Implications of SNOMED CT versioning, *International Journal of Medical Informatics*, 80(6), (2011), 442–453.

Lefèvre, P. : *La recherche d'informations. Du texte intégral au thésaurus*, Hermes Science Publications, 2000.

Lipscomb, C. E. : Medical subject headings (MeSH), *Bulletin of the Medical Library Association*, 88(3), (2000), 265–266.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. : Big data : The next frontier for innovation, competition, and productivity.

Maurer, T. J., & Pierce, H. R. : A comparison of likert scale and traditional measures of self-efficacy, *Journal of Applied Psychology*, 83(2), (1998), 324–329.

Merabti, T. : *Méthodes pour la mise en relations des terminologies médicales : contribution à l'interopérabilité sémantique Inter et Intra terminologique*, Ph.D. thesis, University of Rouen, France, 2010.

Merabti, T., Soualmia, L. F., Grosjean, J., Joubert, M., & Darmoni, S. J. : Aligning biomedical terminologies in french : Towards semantic interoperability in medical applications, in S. Mordechai, & R. Sahu (Eds.), *Medical Informatics*, (pp. 41–68), InTech, 2012.

Merabti, T., Soualmia, L. F., Grosjean, J., Palombi, O., Muller, J.-M., & Darmoni, S. J. : Translating the foundational model of anatomy into french using knowledge-based and lexical methods, *BMC Medical Informatics and Decision Making*, 11(1), (2011), 65.

Merrill, G. H. : The MedDRA paradox, *AMIA Annual Symposium Proceedings*, 2008, (2008), 470–474.

Merrill, G. H. : Concepts and synonymy in the UMLS metathesaurus, *Journal of Biomedical Discovery and Collaboration*, 4, (2009), 7.

Nelson, S. J., Schopen, M., Savage, A. G., Schulman, J.-L., & Arluk, N. : The MeSH translation maintenance system : structure, interface design, and implementation, *Studies in health technology and informatics*, 107(Pt 1), (2004), 67–69.

Noy, N. F., Kunnatur, S., Klein, M., & Musen, M. A. : Tracking changes during ontology evolution, in S. A. McIlraith, D. Plexousakis, & F. v. Harmelen (Eds.), *The Semantic Web – ISWC 2004*, no. 3298 in Lecture Notes in Computer Science, (pp. 259–273), Springer Berlin Heidelberg, 2004.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., & Musen, M. A. : BioPortal : ontologies and integrated data resources at the click of a mouse, *Nucleic Acids Research*, (p. gkp440).

Névéol, A., Grosjean, J., Darmoni, S. J., & Zweigenbaum, P. : Language resources for french in the biomedical domain, in N. C. C. Chair), K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis

(Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland : European Language Resources Association (ELRA), 2014, (pp. 26–31).

Palmer, F. R. : *Semantics*, Cambridge University Press, 1981.

Pathak, J., Solbrig, H. R., Buntrock, J. D., Johnson, T. M., & Chute, C. G. : LexGrid : A framework for representing, storing, and querying biomedical terminologies from simple to sublime, *Journal of the American Medical Informatics Association : JAMIA*, 16(3), (2009), 305–315.

Pereira, S., Névéol, A., Kerdelhue, G., Serrot, E., Joubert, M., & Darmoni, S. J. : Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a french online catalogue, *AMIA Annual Symposium Proceedings*, 2008, (2008), 586–590.

Reis, J. C. D., Pruski, C., Da Silveira, M., & Reynaud-Delaitre, C. : Characterizing semantic mappings adaptation via biomedical KOS evolution : A case study investigating SNOMED CT and ICD, *AMIA Annual Symposium Proceedings*, 2013, (2013), 333–342.

Reis, J. C. D., Pruski, C., Silveira, M. D., & Reynaud, C. : Analyzing and supporting the mapping maintenance problem in biomedical knowledge organization systems, 2012.

Roche, C. : Terminologie et ontologie, *Langages*, 39(157), (2005), 48–62.

Rogers, F. B. : Medical subject headings, *Bulletin of the Medical Library Association*, 51, (1963), 114–116.

Roget, P. M. : *Thesaurus of English Words and Phrases : Classified and Arranged So as to Facilitate the Expression of Ideas and to Assist in Literary Composition*, Gould and Lincoln, 1856.

Rosse, C., & Mejino, J., José L V : A reference ontology for biomedical informatics : the foundational model of anatomy, *Journal of biomedical informatics*, 36(6), (2003), 478–500.

Rzhetsky, A., & Evans, J. A. : War of ontology worlds : Mathematics, computer code, or esperanto ?, *PLoS Comput Biol*, 7(9), (2011), e1002191.

Sakji, S. : *Recherche d'information et indexation automatique des médicaments à l'aide de plusieurs terminologies de santé*, Ph.D. thesis, University of Rouen, France, 2010.

Schulz, S., & Jansen, L. : Formal ontologies in biomedical knowledge representation, *Yearbook of medical informatics*, 8(1), (2013), 132–146.

Schulz, S., Kumar, A., & Bittner, T. : Biomedical ontologies : What part-of is and isn't, *Journal of Biomedical Informatics*, 39(3), (2006), 350–361.

Soualmia, L. F., Sakji, S., Letord, C., Rollin, L., Massari, P., & Darmoni, S. J. : Improving information retrieval with multiple health terminologies in a quality-controlled gateway, *BMC Health Information Science and Systems*, (pp. 1–8).

Soutou, C. : *De UML à SQL : conception de bases de données*, Paris : Eyrolles, 2002.

Stead, W., Hammond, W., & Straube, M. : A chartless record–is it adequate ?, *Proceedings of the Annual Symposium on Computer Application in Medical Care*, (pp. 89–94).

Tardieu, H., Colletti, R., & Rochfeld, A. : *La méthode MERISE. : Démarche et pratiques*, Editions d'Organisation, 1995.

Vandenbussche, P.-Y. : *Définition d'un cadre formel de représentation des Systèmes d'Organisation de la Connaissance*, Ph.D. thesis, Université Pierre et Marie Curie - Paris VI, 2011.

Vanopstal, K., Vander Stichele, R., Laureys, G., & Buysschaert, J. : Vocabularies and retrieval tools in biomedicine : disentangling the terminological knot, *Journal of Medical Systems*, 35(4), (2011), 527–543.

Voelkel, M., Handschuh, S., & Groza, T. : Semantic versioning manager : Integrating SemVersion in protege, 2006.

Wang, Y., Patrick, J., Miller, G., & O'Hallaran, J. : A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT, *BMC Medical Informatics and Decision Making*, 8(Suppl 1), (2008), S5.

Wegner, P. : Interoperability, *ACM Comput. Surv.*, 28(1), (1996), 285–287.

Whetzel, P. L., Noy, N. F., Shah, N. H., Alexander, P. R., Nyulas, C., Tudorache, T., & Musen, M. A. : 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(Web Server issue), (2011), W541–W545.

Yahia, B. B. : Avicenne médecin. sa vie, son œuvre., *Revue d'histoire des sciences et de leurs applications*, 5(4), (1952), 350–358.

Zweigenbaum, P. : Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances, *Innovation Stratégique en Information de Santé*, ISIS(2-3), (1999), 27–47.

Annexes

Annexe A

Illustrations supplémentaires

A.1 Norme ISO 25964-1

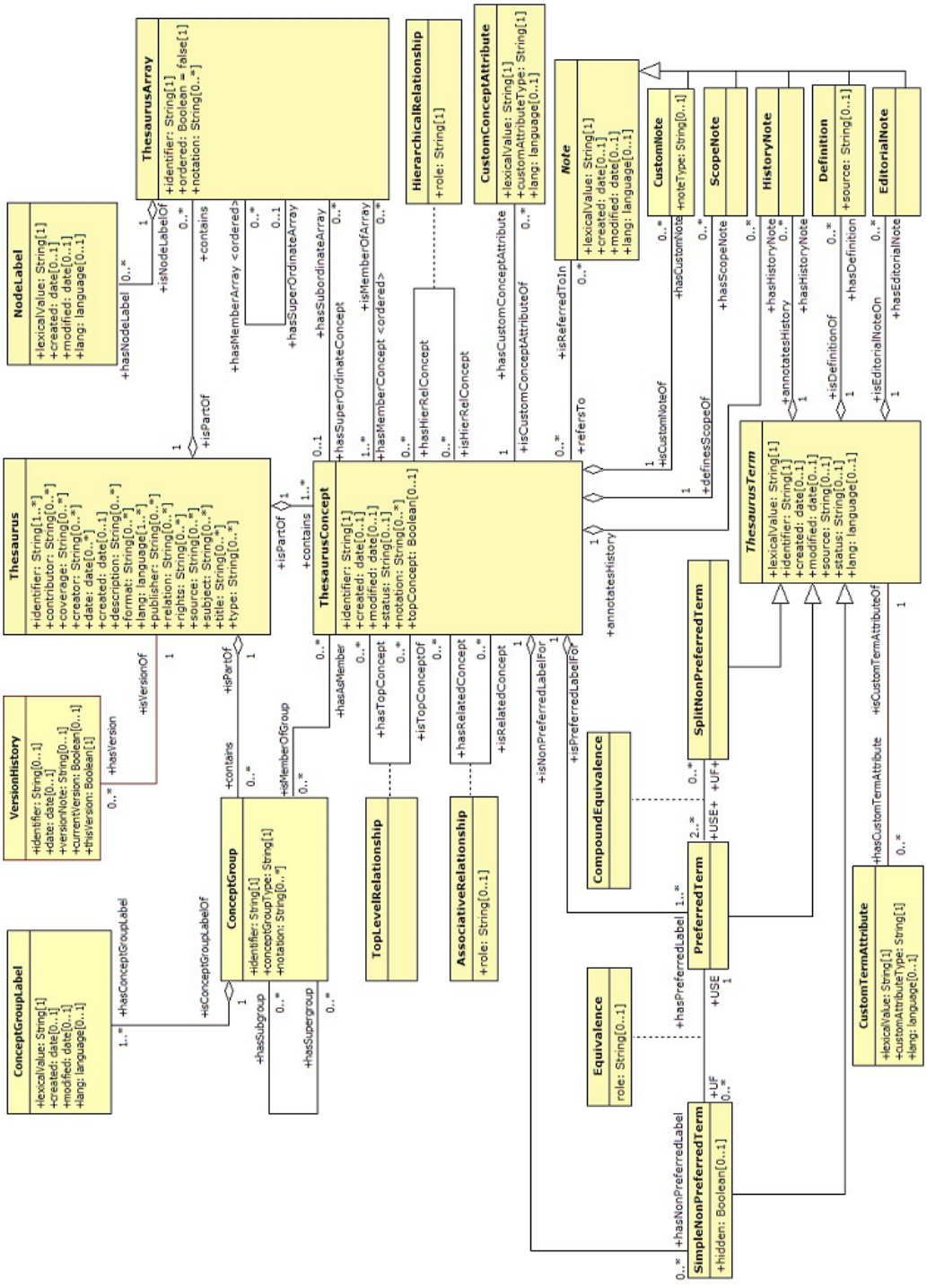

FIGURE A.1 – Modèle de donnée de la norme ISO 25964-1 des thésaurus pour la recherche documentaire

Annexe B

Éléments techniques

B.1 Configurations machines

B.1.1 Machine locale

Intel(r) Core 2 duo E8500 3,16GHz
3 Go de RAM
Disque dur de 300 Go
Windows XP Professionnel SP3

B.1.2 Serveur d'application du parc CISMeF

DELL R710 avec 2 Xeon 5690 (12 Mo de cache, 3,46 GHz), 24 cœurs avec l'hyper-threading
64 Go RAM
Disques SAS SSD (extensibles)
Debian Squeeze (Linux)

Listes des publications

Journaux

Griffon N, Kerdelhué G, Soualmia LF, Merabti T, Grosjean J, Lamy J-B, Venot A, Duclos C, Darmoni SJ. Evaluating alignment quality between iconic language and reference terminologies using similarity metrics. BMC Med Inform Decis Mak. 2014 ;14 :17.

Golbreich C, Grosjean J, Darmoni SJ. The FMA in OWL 2. AIME. Springer-Verlag ; 2011. p. 204-214.

Merabti T, Soualmia LF, Grosjean J, Palombi O, Muller J-M, Darmoni SJ. Translating the Foundational Model of Anatomy into French using knowledge-based and lexical methods. BMC Medical Informatics and Decision Making. 2011 ;11(1) :65.

Conférences internationales à comité de lecture

Grosjean J, Soualmia LF, Bouarech K, Jonquet C, Darmoni SJ. Comparing BioPortal and HeTOP : towards a unique biomedical ontology portal ? 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, 2014.

Grosjean J, Soualmia LF, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Cross-lingual access to biomedical terminologies and ontologies. SWAT4LS (Semantic Web Applications and Tools for Life Sciences) Workshop. 2012.

Grosjean J, Kerdelhué G, Merabti T, Darmoni SJ. The EHTOP : indexing Health resources in a multi-terminology/ontology and cross-lingual world. EAHIL 2012.

Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Teaching medicine with a terminology/ontology portal. MIE. Pisa, Italy ; 2012. p. 949-953.

Grosjean J, Merabti T, Dahamna B, Kergourlay I, Thirion B, Soualmia LF, Darmoni SJ. Health Multi-Terminology Portal : a semantics added-value for patient safety. Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety. 2011. p. 129-138.

Cabot C, Grosjean J, Lelong R, Lefebvre A, Lecroq T, Soualmia LF, Darmoni SJ. Omic Data Modelling for Information Retrieval. Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, 2014.

Golbreich C, Grosjean J, Darmoni SJ. The Foundational Model of Anatomy in OWL 2 and its use. Artif Intell Med. févr 2013 ;57(2) :119-132.

Darmoni SJ, Soualmia LF, Griffon N, Grosjean J, Kerdelhué G, Kergoulay I, Dahamna B. Multi-lingual search engine to access PubMed monolingual subsets : a feasibility study. MEDINFO 2013 - Proceedings of the 14th World Congress on Medical Informatics. 2013.

Merabti T, Soualmia LF, Grosjean J, Letord C, Darmoni SJ. Assisting the translation of SNOMED CT into French. MEDINFO 2013 - Proceedings of the 14th World Congress on Medical Informatics, August 20-23, 2013.

Darmoni SJ, Grosjean J, Merabti T, Griffon N, Dahamna B, Dutoit D. Combining WordNet and Crosslingual multi-terminology health portal to access health information. 6th International Global Wordnet Conference (GWC2012). Matsue, Japan ; 2012. p. 94-99.

Thiessard F, Mougin F, Diallo G, Jouhet V, Cossin S, Garcelon N, et al. RAVEL : Retrieval And Visualization in ELectronic health records. Quality of Life through Quality of Information - Proceedings of MIE 2012. 2012. p. 194-198.

Soualmia LF, Griffon N, Grosjean J, Darmoni SJ. Improving Information Retrieval by Meta-Modelling Medical Terminologies. 13th conference on Artificial Intelligence in MEdicine (AIME). Springer, Heidelberg ; 2011. p. 215-219.

Conférences nationales à comité de lecture

Grosjean J, Merabti T, Darmoni SJ. Health Multi-Terminology Portal. RITS Recherche en Imagerie et Technologie pour la Santé. Rennes ; 2011.

Lelong R, Merabti T, Grosjean J, Joulakian M, Griffon N, Dahamna B, Cuggia M, Pereira S, Grabar N, Thiessard F, Massari P, Darmoni SJ. Moteur de recherche sémantique au sein du dossier du patient informatisé : langage de requêtes spécifique. JFIM 2014, Fès, Maroc.

Darmoni SJ, Soualmia LF, Griffon N, Grosjean J, Kerdelhué G, Kergoulay I, Thirion B, Dahamna B. MLPubMed : une base de données bibliographique multilingue. Colloque RITS. 2013.

Merabti T, Soualmia LF, Grosjean J, Joubert M, Darmoni SJ. Méthodes d'alignement de terminologies médicales et leur intégration dans un portail. IC'2012 :

Atelier IC pour l'intéropérabilité sémantique dans les applications en e-Santé. Paris, France ; 2012.

Dirieh Dibad A, Soualmia LF, Merabti T, Grosjean J, Sakji S, Massari P, Darmoni SJ. Un modèle de données adapté à la recherche d'information dans le dossier patient informatisé : étude, conception et évaluation. Systèmes d'information pour l'amélioration de la qualité en santé Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM). Tunis : Springer ; 2012. p. 251-262.

Merabti T, Grosjean J, Abdoune H, Joubert M, Darmoni SJ. Automatic methods for mapping Biomedical terminologies in a Health Multi-Terminology Portal. EGC 2011 : atelier Extraction de Connaissances et Santé. 2011.

Posters

Grosjean J, Merabti T, Soualmia LF, Letord C, Charlet J, Robinson PN, Darmoni SJ. Integrating the human phenotype ontology into HeTOP terminology-ontology server. Stud Health Technol Inform. 2013 ;192 :961.

Grosjean J, Merabti T, Griffon N, Dahamna B, Soualmia LF, Darmoni SJ. Multiterminology cross-lingual model to create the Health Terminology/Ontology Portal. AMIA. Chicago ; 2012.

Grosjean J, Merabti T, Griffon N, Dahamna B, Darmoni SJ. Multiterminology cross-lingual model to create the European Health Terminology/Ontology Portal. Short papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA. Paris ; 2011. p. 119-122.

Névéol A, Grosjean J, Darmoni SJ, Zweigenbaum P. Language Resources for French in the Biomedical Domain. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland : European Language Resources Association (ELRA) ; 2014. p. 26-31.

Diallo G, Grabar N, Thiessard F, Garcelon N, Grosjean J, Dupuch M, et al. Towards complex queries on data from complex patients. AMIA. Chicago, Illinois ; 2012.

Chapitre de livre

Merabti T, Soualmia LF, Grosjean J, Joubert M, Darmoni SJ. Aligning Biomedical Terminologies in French : Towards Semantic Interoperability in Medical Applications. In : Mordechai S, Sahu R, éditeurs. Medical Informatics . InTech ; 2012. p. 41-68.

Rapports de recherche

Golbreich C, Grosjean J, Darmoni SJ. The Foundational Model of Anatomy in the Health Multi-Terminology Portal. 2011.

Golbreich C, Grosjean J, Darmoni SJ. FMA and HMTP Portal in OWL : Reconciling Ontology with Terminology in Life Sciences via Metamodeling. 2010 juin.

Golbreich C, Grosjean J, Darmoni SJ. Pushing the FMA-OWL Enveloppe Further. CNRS ; 2010.