Normandie Université

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité informatique

Préparée au sein de l'Université Rouen Normandie

Recherche d'information clinomique au sein du Dossier Patient Informatisé : modélisation, implantation et évaluation

Présentée et soutenue par
Chloé CABOT

Thèse soutenue publiquement le 21 décembre 2017 devant le jury composé de		
M. Stéfan Darmoni	PUPH, Université Rouen Normandie	Directeur de thèse
Mme Lorraine Goeuriot	MCU, Université Grenoble Alpes	Examinateur
Mme Marie-Christine Jaulent	DR, LIMICS Inserm U1142	Rapporteur
Mme Lina Soualmia	MCU, HDR, Université Rouen Normandie	Codirectrice de thèse
M. Laurent Vercouter	PU, INSA Rouen Normandie	Examinateur
M. Pierre Zweigenbaum	DR, LIMSI CNRS UPR3251	Rapporteur

Thèse dirigée par :
Pr. Stéfan Darmoni, LITIS EA4108, Université de Rouen Normandie
Dr. Lina Soualmia, LITIS EA4108, Université de Rouen Normandie

UNIVERSITÉ DE ROUEN NORMANDIE

MIIS

litis

Résumé

Les objectifs de cette thèse s'inscrivent dans la large problématique de recherche d'information dans les données issues du Dossier Patient Informatisé (DPI). Les aspects abordés dans cette problématique sont multiples : d'une part la mise en œuvre d'une recherche d'information clinomique au sein du DPI et d'autre part la recherche d'information au sein de données non structurées issues du DPI.

Dans un premier temps, l'un des objectifs de cette thèse est d'intégrer au sein du DPI des informations dépassant le cadre de la médecine pour intégrer des données, informations et connaissances provenant de la biologie moléculaire ; les données omiques, issues de la génomique, protéomique ou encore métabolomique. L'intégration de ce type de données permet d'améliorer les systèmes d'information en santé, leur interopérabilité ainsi que le traitement et l'exploitation des données à des fins cliniques. Un enjeu important est d'assurer l'intégration de données hétérogènes, grâce à des recherches sur les modèles conceptuels de données, sur les ontologies et serveurs terminologiques et sur les entrepôts sémantiques. L'intégration de ces données et leur interprétation selon un même modèle de données conceptuel sont un verrou important. Enfin, il est important d'intégrer recherche clinique et recherche fondamentale afin d'assurer une continuité des connaissances entre recherche et pratique clinique et afin d'appréhender la problématique de personnalisation des soins. Cette thèse aboutit ainsi à la conception et au développement d'un modèle générique des données omiques exploité dans une application prototype de recherche et visualisation dans les données omiques et cliniques d'un échantillon de 2000 patients.

Le second objectif de ma thèse est l'indexation multi terminologique de documents médicaux à travers le développement de l'outil Extracteur de Concepts Multi-Terminologique (ECMT). Il exploite les terminologies intégrées au portail terminologique Health Terminology/Ontology Portal (HeTOP) pour identifier des concepts dans des documents non structurés. Ainsi, à partir d'un document rédigé par un humain, et donc porteur potentiellement d'erreurs de frappe, d'orthographe ou de grammaire, l'enjeu est d'identifier des concepts et ainsi structurer l'information contenue dans le document. Pour la recherche d'information médicale, l'indexation présente un intérêt incontournable pour la recherche dans les documents non structurés, comme les comptes-rendus de séjour ou d'examens. Cette thèse propose plusieurs méthodes et leur évaluation suivant deux axes : l'indexation de textes médicaux à l'aide de plusieurs terminologies et le traitement du langage naturel dans les textes médicaux narratifs.

Abstract

Remerciements

Table des matières

Liste des figures

Liste des tableaux

Liste des acronymes

ADICAP Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologiques. I

ANR Agence Nationale de la Recherche. 3

ANSI American National Standards Institute. 28

ATC Anatomical Therapeutic Chemical classification. I, 93

ATIH Agence Technique de l'Information sur l'Hospitalisation. 23, 93

AUI Atom Unique Identifier. 21

BNCI Base Nationale des Cas d'Intoxications. I

CAC Champ Aléatoire Conditionnel. 39–42

CCAM Classification Commune des Actes Médicaux. I, 44, 67, 92, 93

CDP C Page Dossier Patient. 67

CGH Comparative Genomic Hybridization. 69, 71

CIF Classification Internationale du Fonctionnement, du handicap et de la santé. II

CIM10 Classification Internationale des Maladies, 10 ^{e} révision. ix, II, 23, 24, 32, 44, 67, 91–93, 103, 104, 106–109, 111–113

CLADIMED CLAssification des DIspositifs MÉDicaux. II

CNV Copy Number Variation. 71

CUI Concept Unique Identifier. 21

DPI Dossier Patient Informatisé. 2–4, 6, 27–29, 31, 32, 64, 84

DSM-IV Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision. II

ECMT Extracteur de Concepts Multi Terminologique. 6

FMA Foundational Model of Anatomy. II, 93

GEO Gene Expression Omnibus. 71, 72

GO Gene Ontology. II, 25, 67

HeTOP Health Terminology / Ontology Portal. 4, 80
HGVS Human Genome Variation Society. 71
HL7 Health Level-7. 28
HL7-CCOW HL7 Clinical Context Object Workgroup. 28
HL7-CDA HL7 Clinical Document Architecture. 28
HL7-RIM HL7 Reference Information Model. 28, 32
HPO Human Phenotype Ontology. II, 93
HRDO Human Rare Diseases Ontology. II

i2b2 Informatics for Integrating Biology and the Bedside. 33
ICD-O International Classification of Diseases for Oncology. II
ICNP International Classification for Nursing Practice. III
ICPC-2 International Classification for Patient Safety. II
IDF Inverse Document Frequency. 51, 53
indel Insertion-Délétion. 31
ISO Organisation internationale de normalisation. 28

LOH Loss Of Heterozygosity. 31
LOINC Logical Observation Identifiers Names and Codes. III, 36, 93
LPP Liste des Produits et des Prestations. III
LUI Lexical (term) Unique Identifier. 21

MedDRA MEDical Dictionary for Regulatory Activities terminologies. III, 93
MEDLINE Medical Literature Analysis and Retrieval System Online. 14
MeSH Medical Subject Headings. vii, III, 18–20, 43, 44, 91–93
MVS Machine à vecteurs de support. 40, 42, 43

NCBI National Center for Biotechnology Information. 67, 75
NCIt National Cancer Institute thesaurus. III, 93
NGS Next Generation Sequencing. 29
NIH National Institutes of Health. 68

OMIM Online Mendelian Inheritance in Man. III, 68
OMS Organisation mondiale de la Santé. 93

PHARMA Racines des médicaments. III

PMSI Programme de Médicalisation des Systèmes d'Information. 67

RAVEL Retrieval And Visualization in ELectronic health records. 3

RDF Resource Description Framework. 18

RI Recherche d'Information. 3, 4, 11, 12, 14, 17, 18, 37, 45–48, 50, 51, 53–59, 97

RIDoPI Recherche d'Information dans le Dossier Patient Informatisé. 3

SGBD Système de Gestion de Bases de Données. 64

SIH Système d'Information Hospitalier. 28

SNOMED CT Systematized Nomenclature of Medicine Clinical Terms. III, 24, 25, 32, 41, 44, 93

SNOMED Int. Systematized Nomenclature of Medicine. III

SNP Single Nucleotid Polymorphism. 70, 75, 76

SNV Single Nucleotide Variation. 31

SOC Système d'Organisation des Connaissances. 17, 18, 22, 37

SQL Structured Query Language. 64

SRI Système de Recherche d'Information. 57, 59

SUI String Unique Identifier. 21, 22

SV Structural Variant. 70

TAL Traitement Automatique de la Langue. 35, 37, 39

TecSAN Technologies pour la Santé. 3

TF-IDF Term Frequency-Inverse Document Frequency. 50, 51, 101

TGCA The Genome Cancer Atlas. 68, 72

TSP Thésaurus Santé Publique. III

UCUM The Unified Code for Units of Measure. IV

UMLS Unified Medical Language System. 19–21, 24, 37, 41, 42, 44, 56, 93

XML eXtensible Markup Language. 18, 28, 67

Chapitre 1

Introduction

Sommaire

1.1 Préambule

L'utilisation de données issues de la pratique clinique et de la recherche est un défi majeur dans le domaine de la santé. Dans le domaine numérique, l'intégration de données omiques et de données cliniques dans le Dossier Patient Informatisé (DPI), mais également dans des portails de ressources terminologiques peut permettre le développement de nouveaux tests et thérapies, mais également de comprendre les maladies génétiques complexes, voire des cancers. L'objectif de cette thèse est de proposer un modèle permettant l'intégration de ce type de données complexes, de les intégrer dans le DPI et d'en évaluer l'efficacité dans le contexte de la recherche d'information à grande échelle.

1.2 Contexte de recherche

Les travaux présentés dans la suite de cette section portent sur l'ingénierie des connaissances, et plus particulièrement sur la gestion des vocabulaires contrôlés pour le stockage et l'organisation des connaissances et leur exploitation dans l'indexation de documents et la recherche d'information. Dans ce contexte, plusieurs projets ont été conduits et se poursuivent aujourd'hui. Deux projets sont détaillés ici, les projets PlaIR et RIDoPI, qui sont à l'origine de cette thèse.

1.2.1 Projet PlaIR

PlaIR (Plateforme d'Indexation Régionale) [1] est un projet de recherche cofinancé par l'Union Européenne et la région Haute-Normandie (FEDER pour Fonds Européens DÉveloppement Régional). Coordonné par l'équipe « DocApp » (Document et Apprentissage) du LITIS, PlaIR I s'est déroulé de 2009 à 2012 et avait pour but 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 (voir FIGURE 1.1). 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-terminologies et en multi-disciplines. Il s'agissait donc d'étendre le modèle multi terminologique conçu lors du projet Interopérabilité Sémantique des Terminologies dans les Systèmes d'Information de Santé Français (InterSTIS) et consolidé par l'équipe Traitement de l'Information en Biologie Santé (TIBS) 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 du passage de la mono terminologie à la multi terminologie sont décrits dans la thèse de Julien Grosjean [GROSJEAN, 2014].

Dans PlaIR II, l'objectif est d'améliorer quantitativement et qualitativement le portail de ressources terminologiques largement développé pendant le projet PlaIR I. Le principal axe est d'intégrer dans ce portail des informations dépassant la médecine pour intégrer des données, informations et connaissances provenant de la biologie, ce que nous résumons par les données « omiques » (génomique, protéomique, métabolomique, etc.).

1. http://plair.projets.litislab.fr

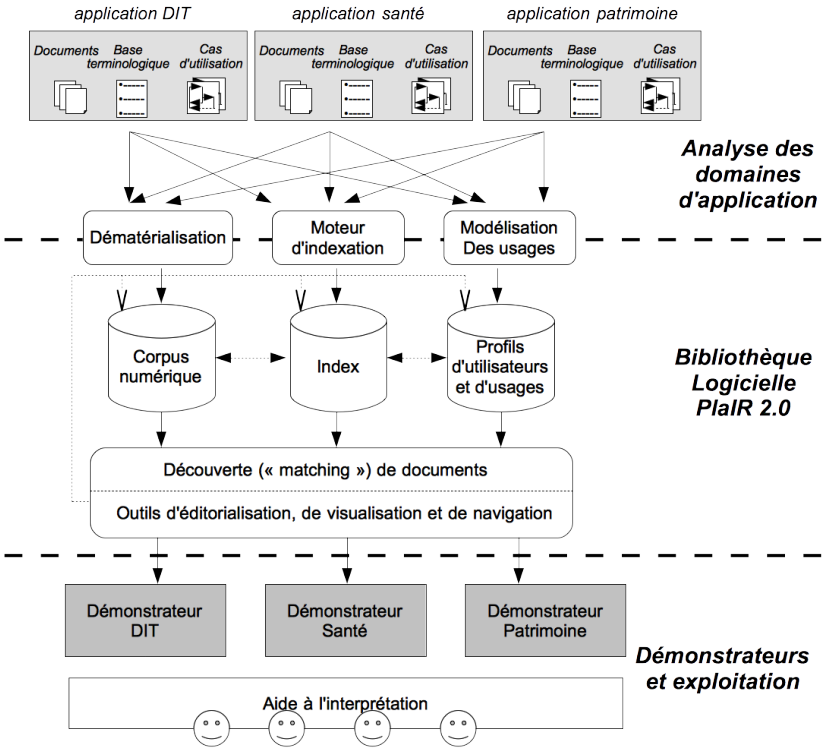

FIGURE 1.1 – Organisation générale de la Plateforme d'Indexation Régionale.

1.2.2 Projet RIDoPI

Depuis 2011, l'équipe TIBS s'intéresse à la problématique de la Recherche d'Information (RI) dans le DPI dans le cadre du projet Recherche d'Information dans le Dossier Patient Informatisé (RIDoPI).

Ce projet a initié un programme de recherche Technologies pour la Santé (TecSAN) de l'Agence Nationale de la Recherche (ANR) (2012–2015) : le projet Retrieval And Visualization in ELectronic health records (RAVEL). Son objectif est de rechercher et d'implémenter des méthodes d'indexation dans le but d'enrichir sémantiquement les données structurées et non structurées au sein du DPI, afin d'optimiser la RI et la visualisation des données cliniques [THIESSARD et al., 2012]. Ce projet réunit des partenaires académiques et industriels. L'équipe TIBS a été en particulier responsable du modèle de données et de la RI.

1.3 Contexte de travail

Dans cette section, je présente l'équipe TIBS, ses travaux de recherche, ainsi que les problématiques et objectifs de mes travaux de thèse.

1.3.1 L'équipe TIBS

L'équipe TIBS (Traitement de l'Information en Biologie Santé, dirigée par le Pr T. Lecroq) est une équipe de recherche rattachée au laboratoire LITIS (Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, EA 4108, dirigé par le Pr T. Paquet).

Le LITIS est le laboratoire de recherche dans le domaine des Sciences et Technologies de l'Information et de la Communication en Haute-Normandie. Le laboratoire a une démarche pluri-disciplinaire, associant l'informatique, les mathématiques, le traitement du signal, la reconnaissance des formes et des images de la médecine. Ses travaux de recherche sont structurés autour de trois axes : l'axe Combinatoire et Algorithmes, l'axe Interaction et systèmes complexes et l'axe Traitement des masses de données. Depuis janvier 2014, le LITIS fait partie de la fédération NormaSTIC, fédération CNRS 3638 avec le Groupe de REcherche en Informatique, Image, Automatique et Instrumentation de Caen (GREYC).

L'un des thèmes généraux des travaux de TIBS est l'ingénierie des connaissances en santé. Par exemple, CISMeF est un projet qui vise à recenser et indexer les ressources (documents, sites, etc.) de qualité en santé disponibles sur l'Internet en langue française [Darmoni et al., 2001]. L'outil Doc'CISMeF initialement développé en 2000 permet de rechercher lesdites ressources. Les travaux de l'équipe s'articulent également autour des terminologies et ontologies de santé, avec le portail Health Terminology / Ontology Portal (HeTOP) qui propose un accès centralisé multilingue aux principales terminologies de santé [Grosjean et al., 2011].

Depuis 2011, l'équipe travaille autour d'un nouvel axe de recherche en informatique biomédicale qui porte sur la RI et la visualisation des données dans le DPI. Ma thèse s'inscrit dans cette contribution.

1.3.2 L'informatique biomédicale

De nombreuses définitions de l'informatique biomédicale se concentrent sur les données, l'information et les connaissances, mais ne fournissent pas une définition adéquate de ces termes. L'informatique biomédicale peut être définie comme la science de l'information appliquée ou étudiée dans le contexte de la biomédecine [Bernstam et al., 2010]. L'American Medical Informatics Association (AMIA) donne une définition plus précise du domaine de l'informatique biomédicale : « L'informatique biomédicale est un domaine interdisciplinaire qui étudie et poursuit l'utilisation efficace des données, des informations et des connaissances biomédicales pour la recherche scientifique, la résolution de problèmes et la prise de décision, motivée par des efforts visant à améliorer la santé humaine. »

Ses activités sont ainsi multiples :

— le développement de théories, méthodes et processus pour la production, le stockage, la récupération, l'utilisation et le partage de données, d'informations et de connaissances biomédicales ;

— l'utilisation des connaissances et technologies de l'informatique, de la communication et plus largement des sciences de l'information et leur application en biomédecine ;

— l'étude, la modélisation, la simulation et l'expérimentation à travers le spectre des molécules aux populations, en traitant une variété de systèmes biologiques, et reliant recherche, pratiques fondamentales et cliniques et systèmes de soins de santé.

Ses caractéristiques placent ainsi l'informatique biomédicale à l'interface de la recherche translationnelle.

1.3.3 La recherche translationnelle

La médecine translationnelle est une discipline qui se développe rapidement, axée sur les technologies et les découvertes en laboratoire et leur traduction dans la pratique clinique. Il existe ici aussi de nombreuses définitions de la médecine translationnelle. L'Institut national de la santé américain (NIH) considère ainsi que « pour améliorer la santé humaine, les découvertes scientifiques doivent être traduites en applications pratiques. Ces découvertes commencent généralement « à la paillasse » avec la recherche fondamentale - dans laquelle les scientifiques étudient la maladie au niveau moléculaire ou cellulaire - puis progressent vers le niveau clinique, ou le « chevet » du patient définissant ainsi la recherche translationnelle comme la recherche *de la paillasse au chevet* » [Zerhouni, 2003]. Plus succinctement, la médecine translationnelle peut être décrite comme la transition de la recherche expérimentale in vitro à l'application chez l'être humain [Wehling, 2015].

Cette transition, ou traduction, nécessite de coordonner et exécuter diverses étapes impliquées dans l'extraction d'informations significatives à partir de données cliniques et expérimentales accumulées de manière ordonnée et semi-automatique :

— la gestion des données haut débit comme le transfert de données, le stockage de données, le contrôle d'accès et la gestion des données ;

— la mise en place d'une infrastructure évolutive ;

— l'analyse de données multidimensionnelles à grande échelle pour l'extraction de connaissances concrètes pour le développement d'applications en biotechnologie et en biomédecine.

1.4 Objectifs

L'un des premiers objectifs de ma thèse est l'intégration au sein du DPI d'informations dépassant le cadre de la médecine pour intégrer des données, informations et connaissances provenant de la biologie moléculaire ; les données omiques, issues de la génomique, protéomique, métabolomique. L'intégration de ce type de données permet d'améliorer les systèmes d'information en santé, leur interopérabilité ainsi que le traitement et l'exploitation des données à des fins cliniques. La modélisation des dossiers, l'exploration des codages et de nouveaux modes de saisies, la sémantique des informations stockées sont alors des éléments essentiels. Un enjeu important est d'assurer l'intégration de données hétérogènes, grâce à des recherches sur les modèles conceptuels de données, sur les ontologies et serveurs terminologiques et sur les entrepôts sémantiques. En France, il n'existe pas de plateforme permettant d'unifier des données médicales (cliniques, biologie ou d'imagerie) de sources hétérogènes et l'utilisation des données du dossier patient. L'intégration de ces données et leur interprétation selon un même modèle de données conceptuel sont un verrou important. L'interconnexion de systèmes d'information et le partage d'information au travers de services web peuvent aider à la transmission ou la non-redondance d'information. Enfin, il est important d'intégrer recherche clinique et recherche fondamentale afin d'assurer une continuité des connaissances entre recherche et pratique clinique et afin d'appréhender la problématique de personnalisation des soins. De nouveaux modèles pour la représentation et l'intégration de connaissances issues à la fois de la médecine et de la biologie moléculaire au sein des dossiers patients sont nécessaires, ainsi que des outils dédiés à la recherche et la visualisation d'information des données cliniques et omiques.

Le second objectif de ma thèse est l'indexation multi terminologique de documents médicaux. Notre équipe développe l'outil ECMT. Il exploite les terminologies intégrées au portail de PlaIR pour identifier des concepts dans des documents non structurés. Ainsi, à partir d'un document rédigé par un humain, et donc porteur potentiellement d'erreurs, de frappe, d'orthographe ou de grammaire, cet outil va identifier des concepts et ainsi structurer l'information contenue dans le document. Pour la recherche d'information médicale, l'indexation présente un intérêt incontournable pour la recherche dans les documents non structurés, comme les comptes-rendus de séjour ou d'examens. Il s'agit donc d'une part d'évaluer l'apport de cet outil à la recherche d'information dans les documents médicaux non structurés contenus dans les DPI et dans un second temps d'améliorer et d'enrichir l'indexation grâce à l'exploitation des réseaux sémantiques et relations inter terminologiques. Les textes médicaux issus de la pratique clinique ayant pour caractéristique de comporter généralement des erreurs qui sont un frein à une indexation efficace, cette problématique doit également être abordée.

1.5 Organisation du mémoire

Ce mémoire s'organise en sept chapitres. Les deux premiers chapitres sont consacrés au contexte de travail, aux objectifs et à l'état de l'art et définitions liés à la recherche d'information en santé. Les deux chapitres suivants présentent les méthodes et modèles utilisés ou conçus ainsi que leur réalisation pour aboutir à une solution d'intégration et de recherche clinomique dans les DPI. Les cinquième et sixième chapitres présentent les travaux réalisés s'intéressant à l'indexation multi-terminologiques de documents médicaux. Le dernier chapitre propose une conclusion de cette thèse et ses perspectives.

Chapitre 2

La recherche d'information en santé

Sommaire

2.1 Introduction

Cette section introduit le domaine d'application de l'informatique médicale. Il vise à contextualiser la revue de la littérature et montrer pourquoi l'informatique médicale est un environnement où la RI a un impact considérable. L'informatique médicale est une discipline à l'interface des sciences de l'information, de l'informatique et de la médecine. Elle traite avec des ressources, appareils et méthodes indispensables à l'acquisition, le stockage, la recherche et l'utilisation d'informatique en médecine clinique et en biologie. Un volume important de ces informations est conservé dans des formats non structurés, notamment en langage naturel. Le langage naturel est répandu pour plusieurs raisons. Les dossiers patients électroniques en sont à leurs balbutiements dans de nombreux pays et ceux ayant implémenté ces projets ont toujours à numériser un volume considérable de données. De plus, alors que les dossiers patients électroniques ont été adoptés suivant différents standards, l'interopérabilité entre ces projets reste une question ouverte. Puis, les professionnels médicaux ont développé des mécanismes de langage naturel sophistiqués et efficients pour communiquer entre eux, par exemple, l'usage intensif d'abréviations et de notations courtes personnalisées. Ainsi, ils peuvent être réticents à remplacer ces usages par l'information structurée nécessaire au traitement automatisé. Le décalage sémantique entre les données médicales brutes, comme les dossiers patients, les données biologiques et la façon dont un humain (un clinicien par exemple) interprète ces données est un problème majeur en informatique médicale [Patel et al., 2007]. L'ambiguïté du langage naturel amplifie ce problème. Les terminologies et ontologies standardisées visent à résoudre cette question en apportant un point de référence sémantique primordial pour l'intégration de données hétérogènes issues de multiples sources. L'accès opportun et pertinent à l'information est essentiel pour le processus de soins. Le défi de gérer cette information implique impérativement un traitement sémantique.

2.1.1 L'information en santé

« L'information » est un concept difficile à définir qui peut être vu de différentes façons. Il est fréquent de voir sa définition à travers la comparaison aux concepts de « donnée » et « connaissance ». Les données consistent en des observations ou mesures réalisées sur l'environnement. L'information représente les données agrégées et organisées décrivant une situation spécifique. La connaissance représente les éléments appris à partir des données et de l'information, cumulés et intégrés au cours du temps, et qui peuvent être appliqués à des faits nouveaux.

Aujourd'hui, nous sommes entrés dans l'ère de la société de l'information [Webster, 2014]. Jamais auparavant une quantité si importante d'information ne fut créée et partagée. Une étude publiée en 2003 estimait que le stockage de nouvelles données

avait augmenté de 30% par an en 1999 et 2002 [Lyman et Varian, 2003]. De fait, l'information est au cœur de notre société et est revenue une ressource indispensable dans chacune de ses sphères. Dans le domaine de la santé, l'information joue un rôle crucial dans les activités professionnelles et les comportements des patients. Deux études de 1966 et 1973 prédisaient que les professionnels de santé passaient environ un tiers de leur temps à gérer et utiliser de l'information [Jydstrup et Gross, 1966; Mamlin et Baker, 1973]. Selon Hersh [Hersh, 2008], il est probable que le temps dédié à la gestion de l'information dans le domaine de la santé soit aussi important, sinon plus, de nos jours.

2.1.2 La recherche d'information dans le domaine de la santé

Probablement pour les mêmes raisons qui ont fait évoluer le domaine de la RI en général au cours des dernières années, la recherche et l'intérêt pour l'application des techniques de ce champ au domaine spécifique de la santé et de la biomédecine ont également connu un réel essor. Le Web et ses applications ont profondément modifié la disponibilité et la facilité d'accès à l'information sur la santé, non seulement pour les professionnels de la santé, mais aussi pour les patients. Pour les professionnels de la santé, les applications offrant un accès facile à des connaissances validées et à jour sur la santé sont d'une grande importance pour la diffusion des connaissances et ont le potentiel d'influer sur la qualité des soins prodigués. D'un autre côté, le Web a ouvert l'accès aux patients, à leur famille et à leur entourage, à des informations sur la santé, à améliorer leurs connaissances et à changer leurs relations avec les professionnels de la santé. Selon Adam Bosworth [Bosworth, 2007], dans un bon système de santé, les patients devraient avoir accès à l'information la plus pertinente possible, ainsi qu'à des services personnalisés de soutien et devraient être capables d'apprendre et d'éduquer ceux souffrant de maux similaires. Pour les professionnels, une des principales et plus anciennes applications en RI est MEDLINE de la National Library of Medicine (NLM) aux États-Unis qui donne accès à la littérature anglophone de recherche biomédicale. Pour les patients, l'information sur la santé est disponible par différents services et avec une qualité différente. Le contrôle et l'accès aux informations sur la santé par les patients restent un sujet sensible, ayant ouvert diverses initiatives gouvernementales partout dans le monde avec l'émergence du domaine de la e-santé. Ce domaine peut être défini comme un domaine émergent à l'intersection de l'informatique médicale et de la santé publique dédié aux services de santé et à l'information fournis via Internet et les technologies qui lui sont liées [Boogerd et al., 2015]. Certaines des principales sociétés privées dans le domaine de la RI, comme Google, Microsoft et Apple ont intensifié leurs investissements dans ce domaine dans le courant des années 2000. Google a lancé en mai 2008 un service intitulé Google Health visant à agréger les données des dossiers patients. Ce service, faiblement utilisé, est définitivement fermé en janvier 2013.

Microsoft a également lancé en octobre 2007 un service intitulé Microsoft HealthVault aux États-Unis qui est étendu aux utilisateurs du Royaume-Uni en 2010. Globalement, le domaine de la e-santé axé sur le patient consiste à donner aux patients plus de pouvoir pour gérer leur santé. Cela peut être fait en leur donnant accès à leurs données de santé, en facilitant l'accès aux informations nécessaires ou encore en fournissant des outils en ligne qui permettent des conseils personnalisés et d'autres approches. Depuis le début des années 2010, le développement des objets connectés, et dans le domaine de la santé, des capteurs biométriques, intégrés à des objets du quotidien a donné naissance à de nombreuses applications de suivi de données biométriques à destination des patients. Apple Health (depuis septembre 2014) ou Google Fit (depuis octobre 2014) propose l'agrégation de ces données. Ces outils permettent le partage de données de santé entre utilisateurs, mais aussi avec des professionnels de santé. Des initiatives ont également été entreprises dans le champ de la recherche clinique (Apple CareKit, Apple ResearchKit) avec le soutien de diverses institutions et entreprises (University of Rochester, Sage Bionetworks, Duke University, University of Cape Town, Johns Hopkins University). Le développement récent de ces outils et leur adoption induit ainsi un développement important des données de santé personnelles, s'ajoutant aux données et ressources disponibles en sciences biomédicales.

2.2 Les ressources en santé et sciences biomédicales

Les informations de santé textuelles peuvent être classées en deux parties : les informations spécifiques aux patients et les informations fondées sur les connaissances [Fagan, 2003; Hersh, 2008]. Le premier type concerne des patients individuels et son but est d'informer les professionnels de la santé de l'état de santé d'un patient. Il comprend généralement le dossier médical du patient qui peut contenir des données structurées ou non (par exemple : résultats de laboratoire, signes vitaux) ou du texte libre (comptes-rendus). Le deuxième type de classification est lié à l'information dérivée et organisée à partir de la recherche observationnelle et expérimentale. Cette information fournit aux professionnels de santé les connaissances acquises dans d'autres situations afin qu'elle puisse être appliquée à des patients individuels ou utilisée pour effectuer d'autres recherches. À l'instar des autres types d'information scientifique, l'information basée sur le savoir peut être subdivisée en informations primaires (résultats directs des recherches originales qui figurent dans des articles, revues ou autres sources) et des informations secondaires (examens, condensations, monographies, documents de revue, directives cliniques, informations sur la santé sur les pages Web et autres sources). Une autre façon de classer l'information basée sur le savoir est de la diviser en quatre sous-catégories : bibliographique, texte intégral, bases de données et agrégations [Fagan, 2003; Hersh, 2008]. Chaque sous-catégorie est décrite ci-après ainsi que certains de

ses principaux exemples.

2.2.1 Ressources bibliographiques

Les ressources bibliographiques en santé sont composées de bases de données bibliographiques, de catalogues en ligne et de registres spécialisés. La distinction entre ces sous-catégories est devenue floue, car, par exemple, les bases de données de référence de la littérature ont commencé à fournir des liens vers la littérature référencée, en se rapprochant des catalogues Web. On peut citer PubMed, Literatura Latino Americana em Ciências da Saúde (LILACS) (pour le portugais et l'espagnol) et Littérature Scientifique en Santé (LiSSa) (équivalent pour le français), développé au sein de notre équipe.

Bases de données de référence de la littérature

Ces bases de données cataloguent des livres et des périodiques et sont les bases de données originales de RI depuis les années 1960, conçues pour guider le chercheur plutôt que fournir directement les ressources. Medical Literature Analysis and Retrieval System Online (MEDLINE) est probablement la base bibliographique la plus connue, issue de la National Library of Medicine (NLM). Dans sa version 2016, elle contient 24 358 442 références à des articles de 5 623 revues dans le domaine de la biomédecine et de la santé. Au cours de l'année 2016, 1 140 078 références ont été ajoutées. MEDLINE est disponible gratuitement via PubMed [1] et une recherche génère une liste de citations (incluant les auteurs, le titre, la source et souvent un résumé) pour les articles de revues, une indication de disponibilité de texte intégral électronique (généralement via PubMed Central [2]) ou un lien vers le site Web de l'éditeur ou tout autre fournisseur de texte intégral. La recherche peut également être réalisée à l'aide de la passerelle NLM [3], une interface Web qui intègre plusieurs systèmes de recherche NLM. D'autres sites donnent également accès à MEDLINE, certains gratuitement et d'autres pour certains frais (généralement en fournissant des services à valeur ajoutée). Outre MEDLINE, la NLM dispose de nombreuses autres bases de données et de ressources électroniques. Leurs bases de données bibliographiques sont organisées en trois catégories : les citations aux revues et autres périodiques depuis 1966 (accessibles par le biais de PubMed qui est composé de MEDLINE, de citations en cours de MEDLINE et de citations fournies par l'éditeur), de citations de livres, de journaux et de matériel audiovisuel disponible auprès de LOCATORplus [4]) et des citations aux articles de revues avant 1966 et des résumés de réunions scientifiques (accessibles via la passerelle NLM).

1. http://pubmed.gov
2. http://www.pubmedcentral.nih.gov
3. http://gateway.nlm.nih.gov
4. http://locatorplus.gov

Outre la NLM, il existe d'autres producteurs de bases de données bibliographiques, tant publiques que privées comme le National Cancer Institute (NCI). Certaines de ces bases de données bibliographiques ont tendance à être plus ciblées sur des ressources ou domaines spécifiques comme la CINAHL (Cumulative Index of Nursing and Allied Health Literature) - la principale base de données non-NLM pour le domaine des soins infirmiers. En France, le Département d'Informatique et d'Informations Médicales du Centre Hospitalier Universitaire de Rouen (D2IM) développe la base de données bibliographique LiSSa [5] contenant plus de 850 000 articles en français [Griffon et al., 2016].

Catalogues Web

Les catalogues Web sont des pages Web qui contiennent des liens vers d'autres pages et sites et partagent de nombreuses fonctionnalités avec des bases de données bibliographiques traditionnelles. Le nombre de ces catalogues augmente [Fagan, 2003]. On peut citer par exemple Doc'CISMeF [6] et HONSelect [7]. Le CHU de Rouen développe, depuis 1995, un catalogue de ressources internet (CISMeF – Catalogue et Index de Sites Médicaux Francophones) dont une des principales priorités est l'enseignement et l'éducation des professionnels de santé et des apprenants (les étudiants en médecine), notamment grâce au recensement de documents pédagogiques. L'outil associé, Doc'CISMeF, permet d'effectuer des recherches dans ce catalogue de ressources, et offre des possibilités d'interrogation plus étendues [Darmoni et al., 2001].

Registres spécialisés

Les registres spécialisés diffèrent des bases de données de référence de la littérature parce qu'ils intègrent des ressources plus diverses. Ce type de ressource d'information peut se chevaucher avec des bases de données de référence de la littérature et des catalogues Web, mais, en général, il pointe vers des ressources d'informations plus diversifiées. On peut citer par exemple le registre de pratiques cliniques de l'ICH (International Council for Harmonisation) [8].

2.2.2 Contenus en texte intégral

Cette sous-catégorie contient des versions en ligne de la version complète des périodiques, des livres et des sites Web. À l'origine, les bases de données en texte intégral étaient principalement des versions en ligne de revues et elles n'ont commencé à inclure des livres qu'avec la baisse du prix des ordinateurs et la croissance du Web.

5. http://www.lissa.fr
6. http://doccismef.chu-rouen.fr/dc/
7. http://www.hon.ch/HONselect/
8. http://www.ich.org/products/guidelines.html

La plupart des périodiques sont aujourd'hui publiés électroniquement. Certains sont diffusés en ligne par l'entreprise responsable de la version imprimée (par exemple Elsevier), d'autres sont exclusivement présents en ligne. Les versions électroniques sont généralement renforcées par des fonctionnalités supplémentaires comme un accès facilité, la fourniture de données supplémentaires telles que des chiffres, des tableaux, des données brutes et des images ou des liens bibliographiques. Seuls certains éditeurs permettent un accès libre et gratuit à leurs revues. Certaines approches très visibles comprennent BiomedCentral (BMC) [9], Public Library of Science (PLOS) [10] et PubMed Central (PMC) [11]. Les manuels éducatifs dans le domaine de la santé sont publiés aussi de plus en plus sur le Web. Ces versions électroniques permettent plusieurs fonctionnalités supplémentaires sur les versions imprimées : elles sont dotées d'images de haute qualité, de contenus multimédias, de liens vers d'autres ressources, de questions d'auto-évaluation interactives et d'un accès plus facile aux mises à jour.

2.2.3 Bases de données

Cette catégorie comprend des bases de données et d'autres catalogues d'informations spécifiques. Ce type d'information est habituellement stocké dans des systèmes de gestion de base de données et contient plusieurs types de ressources comme des images (de radiologie, pathologie et d'autres domaines), des données de biologie moléculaire (séquençage de gène, caractérisation de protéine et d'autres) et des références (qui relient la littérature scientifique). La nature dynamique des bases de données Web les rend plus appropriées à un certain type de contenu. Les images, un élément important de la pratique de la santé, de l'éducation et de la recherche font partie de ces types. Il existe plusieurs bases de données sur les images de santé disponibles sur le Web. L'un des plus célèbres est le Visible Human Project [12], qui consiste en des représentations tridimensionnelles de corps mâles et femelles normaux construits à partir de sections anatomiques. La génomique étudie le matériel génétique dans les organismes vivants et ses recherches ont évolué rapidement ces dernières années. L'un de ses principaux moteurs a été le Human Genome Project, dirigé par le National Human Genome Research Institute lié à l'organisme National Institutes of Health (NIH) [Lander et al., 2001]. Ce projet s'est achevé en avril 2003 avec la production d'une version de la séquence du génome humain qui est disponible gratuitement dans des bases de données publiques [13]. Plusieurs bases de données sur la génomique sont disponibles sur le Web. On peut noter principalement celles publiées par le NCBI aux États-Unis ainsi que celles publiées par l'institut de bio-informatique européen EMBL-EBI. Les ressources

9. http://www.biomedcentral.com
10. http://www.plos.org
11. http://pubmedcentral.gov
12. http://www.nlm.nih.gov/research/visible/visible_human.html
13. https://genome.ucsc.edu/

NCBI sont liées entre elles, ainsi que PubMed via le système Entrez NCBI [14]. Certaines bases de données recensent des données cliniques comme la base Cancer Genome Atlas [15] qui recense les données de biologie moléculaire collectées chez des cohortes de patients atteints de divers cancers ou encore des données liées aux essais cliniques [16].

On peut noter également les bases de données de citations dans la littérature scientifique, impliquée dans l'évaluation de la recherche comme le Science Citation Index (SCI) et le Social Sciences Citation Index (SSCI) de Thomson Reuters, disponible sur le service Web of Science [17].

2.2.4 Agrégations de ressources

Cette dernière catégorie comprend les agrégations des trois premières catégories pour tous les types d'utilisateurs, des patients aux professionnels de la santé et aux scientifiques. La distinction entre cette catégorie et certains des contenus ci-dessus avec plusieurs liens est floue, mais, généralement, les agrégations ont une plus grande variété d'informations qui répondent aux divers besoins de leurs utilisateurs. Il s'agit, par exemple, de sites Web qui collectent plusieurs types de contenus pour générer une ressource cohérente. L'une des plus importantes ressources agrégées d'information est MedlinePlus, un service de la NLM et du NIH mis à jour quotidiennement. Il regroupe des informations provenant de ces entités et d'autres sources de confiance sur plus de 750 pathologies. Il donne également accès à des recherches MEDLINE préformulées explorant des articles de revues médicales, des informations sur les médicaments, une encyclopédie illustrée, un dictionnaire médical, des liens vers des essais cliniques, des didacticiels interactifs et des actualités sur la santé.

L'information en santé est ainsi multiple et conséquente. Son organisation et sa représentation s'avèrent une problématique à part entière détaillée ci-après.

2.3 La représentation des données en santé

Le domaine de la RI met à profit la disponibilité de structures de données bien définies qui peuvent être utilisés dans les processus d'indexation et de RI. L'information sur la santé est de par sa nature très précise et détaillée [Fagan, 2003]. L'organisation de ces connaissances est l'une des plus anciennes applications de la classification, débutée avec les descriptions formelles d'Aristote dans le domaine de la biologie. La représentation des concepts en santé est plus exigeante que dans de nombreux domaines, en raison de ses niveaux de précision, de complexité, de connaissance implicite

14. http://www.ncbi.nlm.nih.gov/Entrez/
15. https://cancergenome.nih.gov/
16. https://clinicaltrials.gov/
17. http://scientific.thomson.com/products/wos/

et d'ampleur des applications. Parallèlement, c'est aussi un domaine riche dans lequel plusieurs systèmes de représentation sont aujourd'hui disponibles. Les Systèmes d'Organisation des Connaissances (SOC) en santé peuvent être classés en trois catégories avec un degré de formalisme croissant : terminologies, thésaurus et ontologies [Vanopstal et al., 2011]. Une terminologie est une liste de termes, qui sont des représentations des concepts utilisés dans un domaine spécifique. Lorsque des relations simples entre des termes différents sont spécifiées, on parle alors de thésaurus. Les relations sont typiquement de trois types : hiérarchiques (les termes sont plus larges ou plus étroits), synonymiques ou associatives (termes avec des relations qui ne sont ni hiérarchiques ni synonymiques). Enfin, les ontologies sont la représentation la plus formelle faisant intervenir des descriptions logiques dans la définition des termes. Les ontologies doivent également présenter une cohérence interne et des sémantiques exploitables de façon automatisée [Gruber, 1993, 1995].

Les SOC en santé décrits dans cette section peuvent être utilisés dans plusieurs domaines de la recherche en informatique médicale, tels que la RI, l'indexation, le traitement du langage naturel, l'interopérabilité sémantique ou encore les systèmes d'aide à la décision. Dans les processus de RI, ils peuvent être utilisés dans le processus d'indexation qu'elle soit manuelle ou automatique et dans le processus de recherche lui-même (par exemple les relations entre concepts peuvent être utilisées pour améliorer l'expression des besoins d'information de la requête).

2.3.1 Représentation des connaissances et vocabulaires contrôlés

Cette section décrit le thésaurus de la NLM qui permet d'indexer la plupart des bases de données de la NLM, suivi d'autres SOC majeurs dans le domaine de la santé.

Medical Subject Headings thesaurus

Le Medical Subject Headings (MeSH) est le thésaurus de la NLM utilisé pour indexer la plupart des bases de données NLM [Coletti et Bleich, 2001]. Il propose des ensembles de termes nommés descripteurs qui sont organisés à la fois en structure alphabétique et hiérarchique et qui permettent la recherche à différents niveaux de spécificité. Les données MeSH peuvent être téléchargées librement sur le site de la NLM, au format eXtensible Markup Language (XML) ou Resource Description Framework (RDF) [18]. Les hiérarchies organisant les descripteurs sont également appelées arbres. Chaque descripteur apparaît dans au moins un arbre et peut apparaître dans autant d'arbres que nécessaire. Le MeSH est structuré en trois niveaux : descripteur, concept et terme. Un descripteur peut être constitué d'une classe de concepts, qui cor-

18. https://www.nlm.nih.gov/mesh/download_mesh.html

```
- Osteoarthritis [C05.799.613] [Descripteur]
    - Osteoarthritis [Concept préféré]
        - Osteoarthrosis [Terme préféré]
        - Arthritis, Degenerative [Terme]
        - Osteoarthrosis Deformans [Terme]
    - Osteoarthrosis Deformans [Concept plus restreint]
        - Osteoarthrosis Deformans [Terme préféré]
```

FIGURE 2.1 – Structure du descripteur *Osteoarthritis* dans le MeSH version 2016 : descripteur, concepts et termes.

respondent à une classe de termes qui sont synonymes les uns des autres. La figure 2.1 donne un exemple d'organisation autour du descripteur *Osteoarthritis* (arthrose). Au niveau hiérarchique le plus élevé, on retrouve des concepts généraux comme *Anatomie* ou *Maladies mentales*. Les concepts plus spécifiques comme *Cheville* ou *Trouble du comportement* sont retrouvés dans la hiérarchie de treize niveaux. Dans sa version 2017, le MeSH comporte 27 883 descripteurs et plus de 87 000 termes permettant de retrouver le concept le plus approprié. Par exemple *Vitamine C* est un terme correspondant à *Acide Ascorbique*. De plus, 232 000 concepts supplémentaires sont disponibles, ciblant principalement les maladies, médicaments et molécules.

Chaque concept a un terme préféré qui est aussi le nom du concept et chaque descripteur a un concept préféré. Le nom du descripteur correspond au terme préféré du concept préféré. De plus, MeSH a deux types de relations : hiérarchique et associative [19]. Le premier type est une composante cruciale d'un thésaurus et est représenté par l'arborescence MeSH qui représente des niveaux distincts de spécificité (termes plus généraux ou plus spécifiques). Les descripteurs MeSH sont organisés en 16 catégories qui peuvent être explorées à travers le navigateur MeSH Browser. Les relations associatives sont souvent représentées par la référence croisée « see related » (voir aussi). Elles peuvent être utilisées pour ajouter ou suggérer des termes à une recherche spécifique, pour signaler dans le thésaurus l'existence d'autres descripteurs qui peuvent être plus appropriés ou pour souligner les distinctions faites dans le thésaurus ou dans la structure hiérarchique du thésaurus.

Outre l'existence de descripteurs, le MeSH a d'autres types de vocabulaires : qualificatifs (ou sous-titres), balises de contrôle, types de publication et concepts supplémentaires. Les qualificatifs peuvent être rattachés aux descripteurs afin de limiter la portée d'un terme (par exemple pharmacothérapie, diagnostic, étiologie, chirurgie). Par exemple, une « déficience en monoamine oxydase » est récupérée par le descripteur *Monoamine Oxydase* combiné avec le qualificatif *Déficient* (« Monoamine oxydase/Déficient »). Il existe des règles limitant la saisie de certains qualificatifs (les qualificatifs admissibles sont mentionnés dans le champ *Qualificateurs Admissibles* pour chaque terme). Les balises de contrôle sont une classe spéciale de descripteurs MeSH

19. http://www.nlm.nih.gov/mesh/meshrels.html

qui doivent être considérés systématiquement pour chaque article, d'où leur nom, et représentent des caractéristiques telles que les espèces, le sexe, l'âge humain, les périodes historiques et la grossesse. Les types de publication décrivent l'élément qui est indexé au lieu de son sujet. Il comporte trois grandes catégories : les composants de la publication (par exemple, résumé), les formats de publication (conférences, articles), les caractéristiques de l'étude (par exemple, essai clinique, méta-analyse). Les fiches descriptives supplémentaires permettent l'indexation des articles avec des descripteurs provenant d'autres thésaurus.

Unified Medical Language System

Le système Unified Medical Language System (UMLS) intègre et distribue des terminologies, classifications et standards et leurs ressources associées. Ce projet a été initié par la NLM, en 1986, du fait de son directeur, Donald Lindberg [Lindberg et al., 1993]. Ce projet visait à réduire les obstacles à l'utilisation des machines dans le domaine de la santé et plus particulièrement à la récupération efficace des informations exploitables par machine [Humphreys et al., 1998; Lindberg et al., 1993]. Deux de ces barrières sont la variété des façons d'exprimer un même concept dans différents vocabulaires et la diffusion d'informations utiles parmi les différents systèmes et leur interopérabilité. De fait, le domaine de l'informatique médicale se caractérise par une grande diversité de vocabulaires développés pour des applications spécifiques (systèmes épidémiologiques, systèmes d'aide à la décision, documentation d'indexation, codes de facturation et procédures). L'absence d'un langage commun freinait l'interopérabilité des applications qui utilisaient ces vocabulaires et était une motivation pour le développement de l'UMLS. L'UMLS se compose de trois sources de connaissances qui peuvent être utilisées séparément ou ensemble. L'un est le Metathesaurus® qui propose plus d'un million de concepts biomédicaux à partir de plus de 100 sources (y compris le MeSH), le deuxième est le Semantic Network® avec 135 types sémantiques et 54 relations sémantiques entre les types, le dernier est le SPECIALIST Lexicon® dans le domaine du traitement automatique de la langue [Kleinsorge et Willis, 2008]. L'utilisation de ces sources de connaissances peut être très diverse (par exemple : recherche d'information, traitement du langage naturel, indexation automatisée, construction de thésaurus, dossiers patients informatisés et autres). Chaque source de connaissances est décrite plus en détail dans cette section. L'UMLS est disponible sous licence libre au téléchargement [20].

Le Metathesaurus® UMLS Le Metathesaurus® est le composant le plus important de l'UMLS. C'est un thésaurus biomédical organisé par concepts, ou significations auquel sont associés des synonymes dans près de 200 autres vocabulaires. Le Metathe-

20. https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

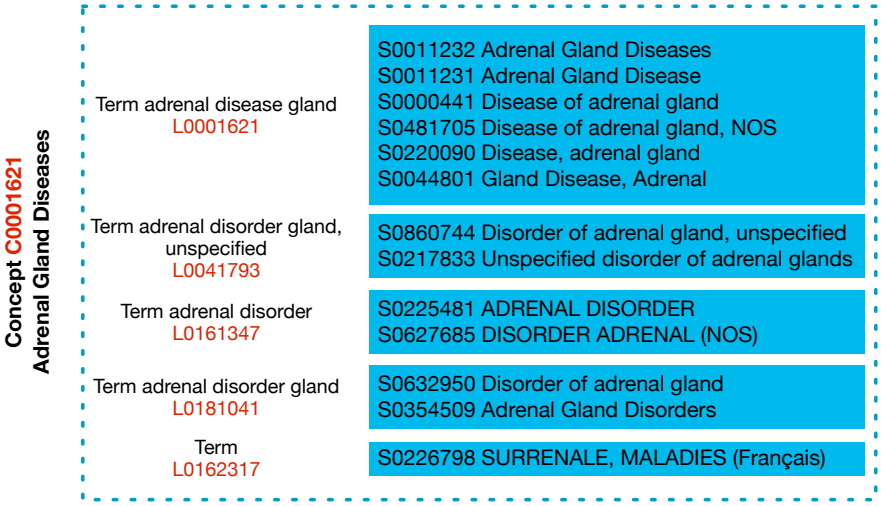

FIGURE 2.2 – La notion de concept au sein du Metathesaurus® UMLS (adapté de KLEINSORGE et WILLIS [2008]).

TABLEAU 2.1 – Métriques générales de la version 2016AB du Metathesaurus® UMLS.

Concepts	3 436 707
Nombre de noms de concepts (AUI)	13 369 382
Nombre de noms de concepts distincts (SUI)	11 287 976
Nombre de noms de concepts distincts normalisés (LUI)	10 291 037
Nombre de sources distinctes intégrées dans le Metathesaurus®	154
Nombre de langages	25

saurus® identifie également les relations entre concepts. Les statistiques concernant les données intégrées dans la dernière version du Metathesaurus® UMLS (version 2016AB) ainsi que les différentes langues gérées sont indiquées dans les tableaux 2.1 et 2.2. Dans le Metathesaurus® les termes synonymes sont regroupés en un concept avec un identifiant unique, le Concept Unique Identifier (CUI) (2.2). Chaque terme, identifié par un identifiant unique, le Lexical (term) Unique Identifier (LUI), est un nom normalisé et peut comporter plusieurs chaînes (identifiées par un String Unique Identifier (SUI)), qui représentent les variantes lexicales dans les vocabulaires sources. Chaque chaîne est associée à un ou plusieurs atomes (identifiés par un Atom Unique Identifier (AUI)) qui représentent le nom du concept dans la source. La figure 2.3 donne l'exemple de la structure du concept *Headache* (CUI :C0018681).

Si deux termes différents ont des significations différentes (par exemple le froid), on leur attribue le même identifiant LUI qui reste associé à différents CUI (par exemple température froide, rhume, sensation de froid). Le même mécanisme peut être reproduit avec des chaînes et des concepts. Le Metathesaurus® est distribué en deux formats : ORF (Original Release Format) et Rich Release Format (RRF). L'accès au Metathesaurus® peut être effectué via le service UMLS Knowledge Service ou le navigateur

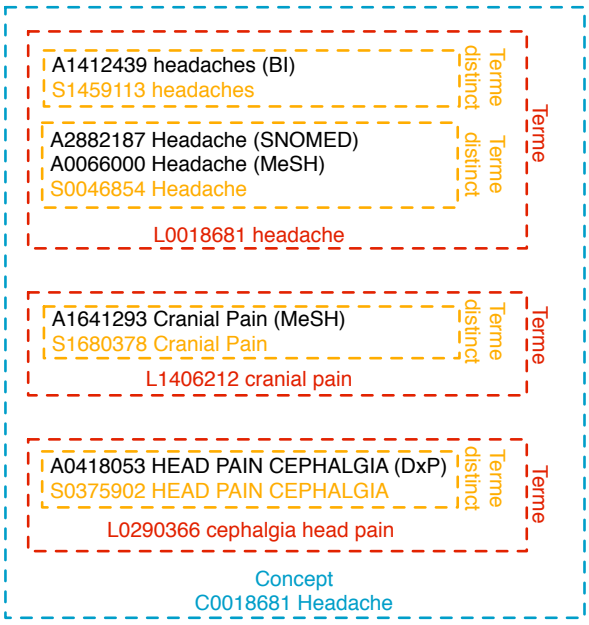

FIGURE 2.3 – Exemple de structure d'un concept avec *C0018681 Headache* (adapté de KLEINSORGE et WILLIS [2008]).

TABLEAU 2.2 – Métriques générales de la version 2016AB du Metathesaurus® UMLS par langue (seules les langues représentant plus de 2% du Metathesaurus® sont ici détaillées).

Langue	Nombre de termes (SUI)	Nombre de ressources	% du Metathesaurus®
Anglais	9 417 453	129	70.44%
Espagnol	1 366 172	9	10.22%
Français	406 771	9	3.04%
Portugais	340 009	5	2.54%
Japonais	314 810	2	2.35%
Néerlandais	280 191	7	2.1%

RRF dans MetamorphoSys.

Semantic Network® Le Semantic Network® est une ontologie de haut niveau dans le domaine de la santé [CHEN et al., 2006], composée de 135 types sémantiques, qui peuvent être assignés aux concepts du Metathesaurus® et 54 relations sémantiques, un ensemble de relations associées aux types sémantiques. Les types sémantiques sont les nœuds du réseau et les relations sont les arcs. Il est fourni dans un format de table relationnelle et dans un format d'enregistrement unitaire. Les types sémantiques sont organisés en deux hiérarchies : *Entité* et *Évènement* et sa portée actuelle est très large, permettant la catégorisation sémantique d'une large gamme de SOC. Chaque concept du Metathesaurus® est associé à au moins un type sémantique (le type le plus spécifique disponible dans la hiérarchie). Au lieu d'ajouter des types sémantiques au réseau pour englober un objet dans les catégories les plus appropriées, les concepts

qui n'appartiennent pas à un niveau de granularité doivent être associés à un type d'un niveau supérieur. Par exemple, le type sémantique *Objet fabriqué* a deux nœuds enfants : *Dispositif médical* et *Dispositif de recherche*. Si un objet n'est ni un dispositif médical ni un dispositif de recherche, il est simplement affecté au type plus général *Objet fabriqué*. Les relations sémantiques peuvent être hiérarchiques ou associatives. Le lien *is a* (*est un*) est le lien principal dans le réseau qui établit la hiérarchie des types et des relations (par exemple, l'animal est un organisme). L'ensemble des associations sont regroupées en cinq grandes catégories : « physiquement liées à », « liées spatialement à », « temporellement liées à », « fonctionnellement liées à » et « conceptuellement liés à ». Dans la mesure du possible, les relations sont définies entre les types sémantiques de plus haut niveau et, en général, sont héritées par tous les descendants de ces types. Les relations ne s'appliquent pas nécessairement à toutes les instances de concepts qui ont été affectées aux types sémantiques qui sont les nœuds de ce lien. Si cela n'a aucun sens, l'héritage des relations peut également être bloqué à un seul ou à tous les descendants des types sémantiques qui sont des liens.

SPECIALIST LEXICON SPECIALIST LEXICON a deux composantes principales : le lexique et ses outils. Le lexique est un lexique anglais général de mots communs qui comprend de nombreux termes biomédicaux et a été développé pour le support du traitement automatique de la langue. Les outils sont des programmes qui traitent les termes. Les entrées du SPECIALIST LEXICON enregistrent la syntaxe, la morphologie (inflexion, dérivation et composition) et l'orthographe de chaque terme. Les éléments lexicaux peuvent être composés de plus d'un terme s'il s'agit d'une expansion d'acronymes et d'abréviations généralement utilisés.

La Classification Internationale des Maladies - 10 e révision

La CIM10 est une liste de classification médicale de l'Organisation mondiale de la santé (OMS). Elle contient des codes pour les maladies, les signes et les symptômes, les découvertes anormales, les plaintes, les circonstances sociales et les causes externes de blessures ou de maladies. La table analytique comporte vingt-deux chapitres depuis 2006, du fait de sa plus récente mise à jour ; elle en comptait vingt et un auparavant. Chaque chapitre est divisé en catégories affectées d'un code à trois caractères, par exemple : asthme J45 (voir Tableau 2.3). La majorité des catégories propose un niveau de détail supplémentaire ou sous-catégorie dont le code est précisé par un quatrième caractère séparé des trois premiers par un point (par exemple : asthme allergique J45.0). Le nombre de codes utilisables de la CIM10 incluant les extensions de circonstances et de lieux du chapitre XX est de 16 800.

L'OMS fournit des informations détaillées en ligne et met à disposition un ensemble de documents. La version internationale de la CIM10 ne doit pas être confondue avec les

TABLEAU 2.3 – Exemples de catégories incluses dans la CIM10.

Code de la catégorie	Nom de la catégorie	Exemples de maladies
J11	grippe, virus non identifié	grippe
J00	rhinopharyngite aiguë [rhume banal]	rhume
J98	autres troubles respiratoires	emphysème compensateur
J93	pneumothorax	pneumothorax spontané

modifications nationales de la CIM10 qui comportent souvent beaucoup plus de détails et ont parfois des sections distinctes pour les procédures. La Modification clinique de la CIM10 aux États-Unis (ICD-10-CM), par exemple, comporte environ 68 000 codes. Les États-Unis disposent également du Système de codification des procédures (ICD-10-PCS) et un système de codage qui contient 76 000 codes de procédure qui ne sont pas utilisés par d'autres pays. En France, l'Agence Technique de l'Information sur l'Hospitalisation (ATIH) édite chaque année une version actualisée complète du volume 1 (Table analytique) de la CIM10 en collaboration avec l'OMS. Ce document intègre les mises à jour de l'OMS et les extensions et modifications réalisées par l'ATIH. Cette version comporte 17 300 codes utilisables. Après des versions alpha et bêta soumises au public dès juillet 2011 pour la première puis mai 2012 pour la seconde, la version consolidée de la CIM-11, version ontologique de la CIM, doit être soumise à l'Assemblée mondiale de la santé dès mai 2018 pour sa commercialisation officielle.

Systematized Nomenclature of Medicine Clinical Terms

La Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT), en français Nomenclature Systématisée de la Médecine - Termes cliniques, est une ontologie incluse dans l'UMLS. Il s'agit d'une collection de termes médicaux couvrant une large gamme de concepts, y compris dont les pathologies, procédures, organismes, structures corporelles et produits pharmaceutiques [SPACKMAN, 2008]. La SNOMED CT est l'une des plus importantes ontologies spécifiques au domaine de la santé, avec environ 326 734 concepts, 1 299 729 termes et 1 672 322 relations. La SNOMED CT utilise les logiques de description comme une représentation formelle sous-jacente, de sorte qu'il s'agit strictement d'une représentation symbolique des connaissances [FRIXIONE et LIETO, 2012].

Les concepts de SNOMED CT sont représentés comme des nœuds dans un graphe acyclique. Chaque concept possède un identifiant unique et un certain nombre de descriptions alternatives pour ce concept. Les concepts peuvent être divisés en plusieurs catégories de haut niveau dont la répartition se trouve dans la FIGURE 2.4. Les concepts de SNOMED CT peuvent être définis en termes de relations avec d'autres concepts. La relation la plus fondamentale est l'héritage, ou parent-enfant. Ainsi, les concepts sont organisés en une hiérarchie héréditaire. Par exemple, la FIGURE 2.5 montre le concept

Pneumonie virale comme un enfant de *Pneumonie infectieuse*. En plus des héritages, un certain nombre d'autres relations peuvent être définies entre les concepts. La figure montre que le concept *Pneumonie virale* a une relation *agent causal* avec le concept *Virus*.

La SNOMED CT couvre un large éventail de connaissances médicales dans une seule ressource autonome, alors que l'UMLS est en fait un conglomérat de ressources différentes, chacune avec une couverture variable. En outre, la SNOMED CT a un processus rigoureux de contrôle qualité supervisé par l'Organisation internationale de développement de la santé.

La SNOMED CT permet la construction d'expressions post-coordonnées. La post-coordination permet aux utilisateurs de spécifier un nouveau concept en combinant plusieurs concepts SNOMED CT. Par exemple, une expression pour décrire un acide aminé hydrophile essentiel pourrait être composée des concepts *acide aminé hydrophile* et *acide aminé essentiel*. À leur tour, ils peuvent être composés avec d'autres concepts qualifiant la nature d'un acide aminé. Dans cette approche, un raisonneur peut être utilisé pour déterminer la relation entre les expressions construites à la volée et les classes de l'ontologie de base, c'est-à-dire que les expressions peuvent être classées et placées à l'emplacement correct dans la hiérarchie des concepts [Dhombres et al., 2015; Karlsson et al., 2014]. L'utilisation d'expressions post-coordonnées dans les systèmes d'information nécessite de préciser les relations exactes entre les expressions post-coordonnées et le contenu SNOMED CT existant ainsi que de respecter les contraintes définies par le modèle de la SNOMED CT.

Gene Ontology

Le projet Gene Ontology (GO) est un effort collaboratif pour répondre à la nécessité d'une description cohérente des produits génétiques dans les bases de données. Fondée en 1998, le projet a commencé comme une collaboration entre trois bases de données d'organismes modèles, FlyBase (Drosophila), la base de données génomique Saccharomyces (SGD) et la base de données du génome de souris (MGD). Le projet GO a développé trois ontologies structurées qui décrivent les produits génétiques en termes de processus biologiques, de composants cellulaires et de fonctions moléculaires associés de manière indépendante des espèces. Leur répartition est décrite dans la Figure 2.6.

L'utilisation des termes GO en collaboration avec des bases de données facilite des requêtes uniformes dans tous les domaines. Les vocabulaires contrôlés sont structurés afin qu'ils puissent être interrogés à différents niveaux. Par exemple, les utilisateurs peuvent demander à trouver tous les gènes dans le génome de la souris qui sont impliqués dans la transduction du signal, ou de zoomer sur toutes les tyrosine kinases réceptrices qui ont été annotées. Cette structure permet également aux annotateurs d'attribuer des propriétés aux gènes ou aux produits de gènes à différents niveaux, en

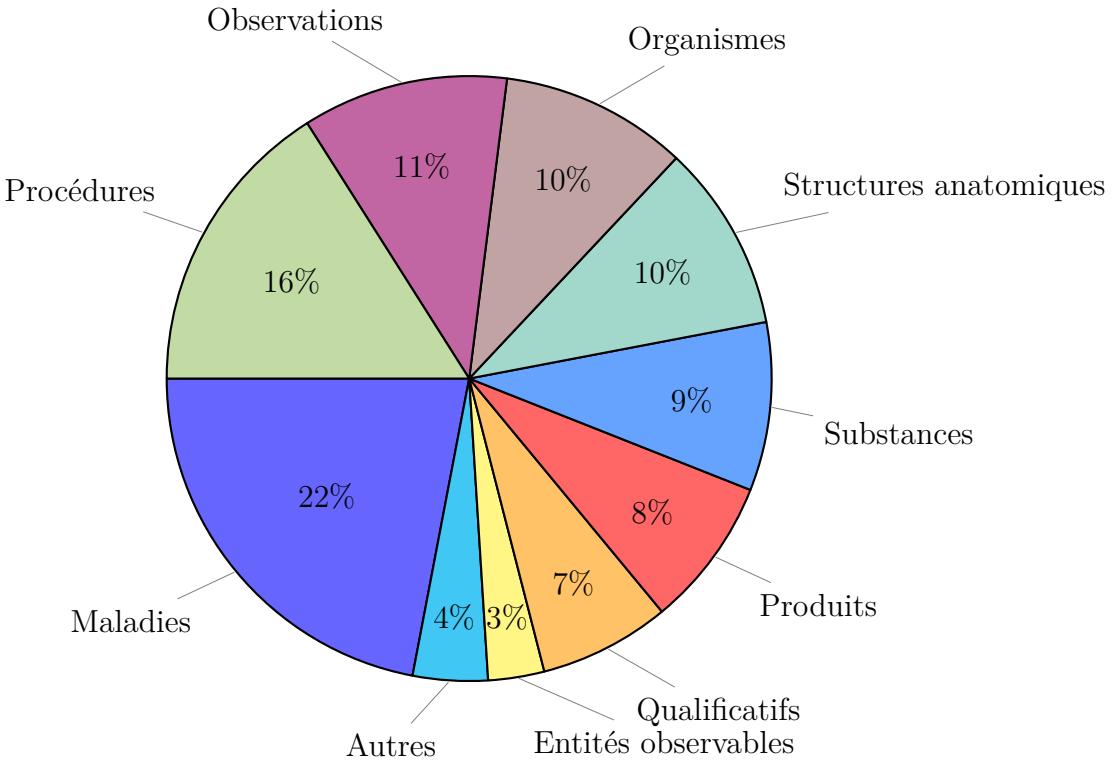

FIGURE 2.4 – Répartition des catégories de concepts dans l'ontologie SNOMED CT.

fonction des connaissances sur cette entité.

2.3.2 Données cliniques et Dossiers Patients Informatisés

Définitions

Depuis l'émergence du concept d'un dossier de santé interopérable, exhaustif et axé sur le patient dans les années 90, les différentes acceptions d'un tel dossier ont toujours été motivées par l'idée de soutenir les soins de santé et de maintenir, respectivement, améliorer, sa qualité [WAEGEMANN, 2002]. Bien que cette idée de base ait perduré, les éléments spécifiques contenus ou le nom qui a été donné à ces différents concepts ont souvent changé avec le temps [WAEGEMANN, 2003]. À l'heure actuelle en France, le terme DPI est largement utilisé. Il décrit le concept d'une collecte globale et transversale des données sur la santé et les soins de santé d'un patient. Il comprend donc des données qui ne sont pas seulement particulièrement pertinentes pour le traitement médical d'un sujet, mais aussi pour la santé d'un sujet en général. Le patient est considéré comme un partenaire actif dans son traitement en accédant, en ajoutant et en gérant les données liées à la santé, en soutenant ainsi les soins [BALL et al., 2007]. Indépendamment de la revendication d'améliorer la qualité et de soutenir les soins de santé, de nouveaux défis ont surgi avec le temps et qui devront être abordés par des DPI modernes. On peut notamment mentionner le coût qui est devenu un facteur critique dans les soins

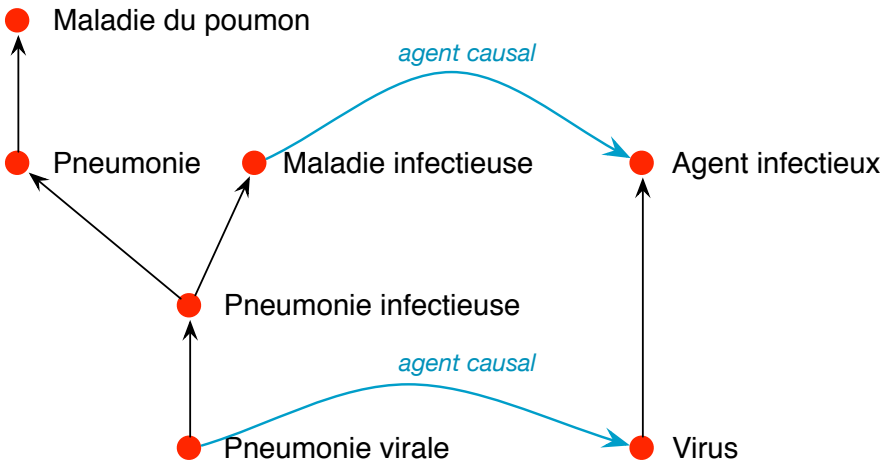

FIGURE 2.5 – Hiérarchie du concept *Pneumonie virale*.

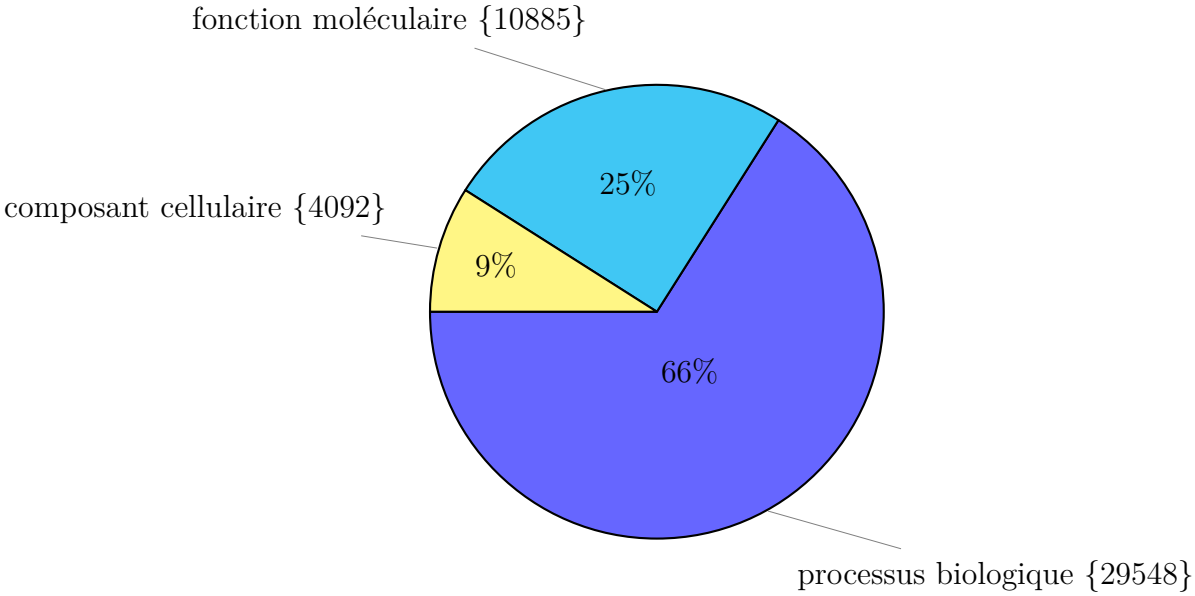

FIGURE 2.6 – Répartition des termes dans les trois ontologies de la Gene Ontology.

de santé et, par conséquent, a également une forte influence sur le développement des DPI [IAKOVIDIS, 1998]. Un autre problème concernant les DPI est la nécessité d'une interopérabilité sur un plan technique, syntaxique [GRIFFON, 2013] mais aussi sémantique [MARY, 2017]. Cette interopérabilité peut se concevoir à différentes échelles, non seulement à l'échelle nationale entre les différents acteurs de santé, mais aussi transfrontalière avec la mobilité professionnelle et privée croissante au sein de l'Union Européenne [HOERBST et AMMENWERTH, 2010]. Enfin, la confidentialité et la sécurité des données contenues dans les DPI demeurent une préoccupation constante.

Caractéristiques des DPI

Contenu Les DPI visent à regrouper toute information clinique, voire administrative, permettant la prise en charge du patient. Ces informations peuvent inclure : les informations personnelles, les actes médicaux et hospitalisations, les antécédents personnels et familiaux, la liste des professionnels de santé impliqués dans le suivi, les allergies, les données de surveillance collectées par le patient, les recommandations, les vaccinations, les prescriptions, les notes cliniques et enfin les résultats d'examen biologique ou d'imagerie [Archer et al., 2011]. Les informations fournies par les praticiens devraient employer un vocabulaire facile d'accès pour les non-praticiens. Parallèlement, les informations saisies par les patients peuvent ne pas être aussi complètes, précises et organisées que les données échangées entre les acteurs de soins de santé. Le contenu doit être important, compréhensible et crédible pour les patients et leurs soignants et approprié pour l'accès au Web par des personnes autorisées par un patient. L'expérience des médecins a montré que les listes de maladies et diagnostics, les notes cliniques, les médicaments et les données sur les allergies et les résultats des tests de laboratoire et de diagnostic peuvent être partagés avec les patients [Halamka et al., 2008].

Architecture L'interopérabilité entre les systèmes est essentielle pour le partage des données entre les différents acteurs : patient, hôpital, médecin traitant et autres praticiens et les laboratoires. Depuis les années 90, des efforts de standardisation ont été entrepris, notamment avec le développement de la version 3 de Health Level-7 (HL7) Stolyar et al. [2005]. HL7 est un ensemble de normes et spécifications techniques pour les échanges informatisés de données cliniques, financières et administratives entre les Système d'Information Hospitalier (SIH). Ces spécifications sont intégrées au corpus des normes formelles américaines (American National Standards Institute (ANSI)) et internationales (Organisation internationale de normalisation (ISO)). HL7 développe des standards conceptuels (HL7 Reference Information Model (HL7-RIM)), des standards de documents (HL7 Clinical Document Architecture (HL7-CDA)), des standards concernant les applications (HL7 Clinical Context Object Workgroup (HL7-CCOW)) et enfin des standards concernant les échanges d'information (HL7 v2.x puis HL7 v3.0). Initialement américaines, ces spécifications s'exportent et tendent à devenir un standard international. Elles définissent structure et rôle des messages pour permettre une communication efficace des données liées au système de santé. Par exemple, HL7-RIM propose une représentation du domaine clinique HL7 et identifie le cycle de vie des messages ou de groupes de messages. C'est un modèle partagé entre tous les domaines HL7. HL7-CDA est un standard XML spécifiant la structure et la sémantique des documents cliniques pour leur échange entre systèmes.

Confidentialité et sécurité Les deux tiers des patients adultes s'inquiètent de la vie privée et de la sécurité de leurs informations sur la santé, mais la plupart de ceux qui utilisent des DPI ne s'inquiètent pas des implications pour la vie privée [California

HealthCare Foundation, 2010].

Avantages et limites à l'utilisation des DPI

Avantages De nombreux travaux ont étudié les bénéfices des DPI en considérant les résultats cliniques, organisationnels et sociétaux. Les résultats cliniques comprennent l'amélioration de la qualité des soins, une réduction des erreurs médicales et d'autres améliorations dans les indicateurs qui décrivent la pertinence des soins. Les résultats organisationnels, d'autre part, ont inclus des éléments tels que la performance financière et opérationnelle, ainsi que la satisfaction chez les patients et les cliniciens qui utilisent les DPI. Enfin, les résultats au niveau sociétal comprennent l'amélioration de la recherche et l'amélioration de la santé de la population.

Au niveau clinique, les travaux sont particulièrement axés sur la qualité des soins et la sécurité des patients. Les DPI ont été empiriquement liés à une adhésion accrue aux directives cliniques et recommandations [Dexter et al., 2009]. Un autre bénéfice moins tangible associé aux DPI est la capacité améliorée à mener des recherches. Le fait de disposer de données sur les patients stockées électroniquement augmente la disponibilité des données ce qui rend possible et facilite de nombreuses tâches en recherche clinique.

Limites L'adoption des DPI présente de nombreux obstacles perçus et réels. Tout comme pour toute nouvelle technologie, l'échec peut souvent être lié à une faible implication des acteurs lors des phases de planification, de la conception et de la mise en œuvre. Le manque de confiance est un autre obstacle, de même que les mauvaises connaissances en informatique ou une accessibilité inadéquate. Le DPI idéal semble être celui qui donne accès à la totalité ou à la plupart des informations cliniques du patient [Detmer et al., 2008; Tang et al., 2006]. Cela nécessite que l'information du patient soit intégrée à travers des réseaux interopérables qui recueillent des informations provenant d'installations qui ont traité le patient. Ces DPI sont intégrés au système de santé. Il existe un certain nombre d'obstacles techniques et non techniques à la réussite de la mise en œuvre de ces DPI idéaux présentés dans le Tableau 2.4 [Archer et al., 2011].

Aujourd'hui, les DPI n'incluent pas de manière standardisée des données clinomiques comme des données de séquençage ou d'expression protéique. Ces données sont pourtant centrales en recherche translationnelle.

2.3.3 Données biologiques et recherche translationnelle

Ainsi, depuis le séquençage du génome humain en 2001, les technologies de séquençage haut débit ou Next Generation Sequencing (NGS) sont devenues peu à peu accessibles et remplacent progressivement dans les laboratoires les techniques basées

TABLEAU 2.4 – Barrières à l'adoption, au déploiement et à l'utilisation des DPI (adapté de ARCHER et al. [2011]).

Limites	Implications
Systèmes d'information de santé	— Équilibre entre l'autonomie du médecin et du patient — Manque de formation technologique, d'intérêt ou de capacité des médecins — Résistance au changement — Portée du travail et des responsabilités des prestataires de soins de santé — Rémunération et incitation des médecins — Préoccupations du fournisseur concernant les risques de responsabilité
Confiance des acteurs	— Préserver la confidentialité des informations médicales
Normes techniques pour l'interopérabilité des systèmes	— Normes d'échange de données — Normes minimales de définition de données dans les spécialisations spécifiques de fournisseur — Normes de sécurité et de confidentialité — Certification des produits de technologie de l'information sur la santé
Manque d'adoption par les praticiens, les institutions	— En Europe et en Amérique du Nord
Manque d'infrastructures de technologie de l'information	— Manque de ressources prenant en charge l'intégration du système — Étendue des systèmes non compatibles existants — Besoin de médiation des réseaux, structures organisationnelles pour soutenir l'intégration — Services en ligne limités chez les fournisseurs de soins de santé et les institutions
Fracture numérique	— Considérations relatives à la situation scolaire et socioéconomique — Connaissances en santé — Besoins spéciaux : limitations visuelles, cognitives ou physiques — Ressources financières

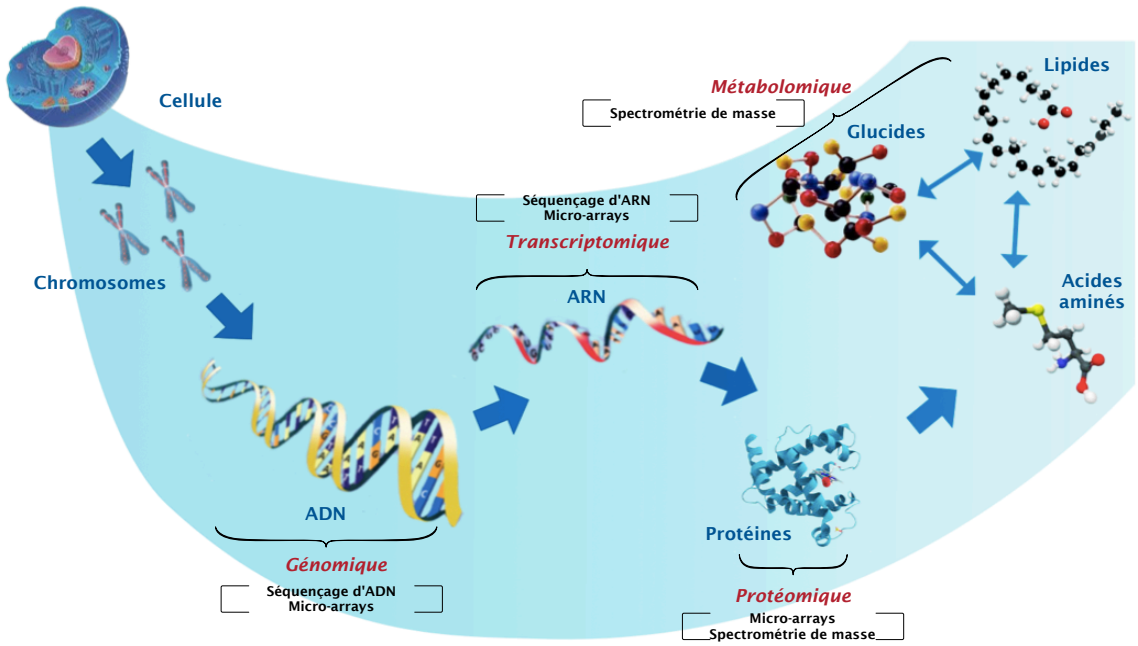

FIGURE 2.7 – Les principales sciences omiques et leurs méthodes d'étude.

sur les puces [NG et al., 2010]. Alors que le séquençage du génome humain avait pris 13 ans et coûté plusieurs milliards de dollars, aujourd'hui un laboratoire peut séquencer un génome complet en une semaine pour quelques milliers de dollars, et bientôt pour moins de 1000 dollars [FERNALD et al., 2011]. Les techniques NGS se distinguent également sur des points clés : quantité de matériel biologique de départ nécessaire, personnel requis et taux d'erreurs moindre. Elles sont aujourd'hui utilisées pour répondre à de nombreuses questions biologiques à l'échelle d'un génome : détermination de variations génomiques (DNA-Seq), inventaire de transcrits et analyse de leur expression (RNA-Seq), interaction ADN-protéines ou encore modifications de la chromatine pour la régulation de l'expression des gènes (ChIP-Seq/Méthyl-Seq).

Le développement de ces méthodes a permis l'essor de nouvelles disciplines dédiées à l'étude de composants biologiques tels que les gènes avec la génomique, les protéines avec la protéomique, les métabolites avec la métabolomique. Par analogie avec le mot génome, leur nom est formé du suffixe *-omique*, un néologisme désignant un ensemble, une collection : l'*ome* (voir FIGURE 2.7). Ainsi la gén*omique* signifie l'étude du génome, l'ensemble du matériel génétique d'un individu.

L'étude du génome fournit les clés de la compréhension de fonctions biologiques de l'individu grâce à l'étude de certaines anomalies : Single Nucleotide Variation (SNV), Insertion-Délétion (indel), translocations, changement de ploïdie, perte d'hétérozygotie (Loss Of Heterozygosity (LOH)). L'analyse des séquences peut offrir des possibilités supplémentaires pour le diagnostic, et par extension pour la génomique personnelle qui vise à identifier les caractéristiques génétiques d'un individu, évaluer ses risques de développer une maladie, et adapter les soins à ces caractéristiques.

D'autres disciplines -*omiques* peuvent révéler une grande utilité dans le domaine de la santé. Par exemple, l'épigénomique informe sur les états d'activités physiologiques ou pathologiques dans la cellule. Des modifications pathogènes peuvent être la marque de certains cancers ou pathologies.

La masse de données omiques générée par l'utilisation croissante de ces méthodes ouvre ainsi de nouvelles voies dans la recherche d'applications biomédicales. En effet, ces nouvelles techniques peuvent être utilisées dans la démarche clinique, notamment pour élaborer des tests diagnostiques. Elles peuvent être également utiles dans la prise en charge du patient, grâce à l'élaboration de nouvelles thérapies (pharmacogénomique). En recherche clinique, l'analyse des séquences permet de comprendre certaines pathologies, en particulier les maladies génétiques complexes ou encore certains cancers [Sarkar et al., 2011]. Enfin, l'intégration des données omiques aux données du patient peut également répondre à de nombreuses questions épidémiologiques. De nouvelles thérapies et de nouvelles méthodes de prévention sont actuellement discutées et développées. Ces approches innovantes ne peuvent être poursuivies sans gérer les vastes quantités de données générées dans les laboratoires dans les domaines tels que la génomique fonctionnelle ou la protéomique. Dans cet environnement, le DPI, des systèmes d'aide à la décision ou encore des techniques de traitement du signal peuvent être intégrés pour élargir le champ de l'information.

La recherche translationnelle est la discipline émergente qui vise à créer cet espace d'information commun qui pourrait mener à la découverte de nouvelles méthodes thérapeutiques et de diagnostic [Sarkar et al., 2011], à l'interface de la biologie et la médecine. L'utilisation des connaissances produites par la recherche scientifique dans le domaine de la santé est l'objectif principal de la médecine translationnelle. En effet, la médecine translationnelle s'intéresse à l'application de résultats de recherche en biologie, et en particulier des résultats des études *omiques* (et on parlera alors de bioinformatique translationnelle), dans le but de faire converger la recherche et le patient. Dans ce but, le développement de plateformes visant à la gestion et à l'exploitation des données produites est indispensable.

2.4 Entrepôts de données et plateformes de recherche biomédicales

Actuellement, peu de systèmes d'information hospitaliers permettent d'intégrer ces données aux données cliniques au sein des DPI [Aronson et Rehm, 2015]. En revanche, depuis 2010, plusieurs projets de plateformes de recherche translationnelle ont été initiés, visant à la réutilisation des données biomédicales et à l'exploitation des connaissances fondamentales. En Europe par exemple, le projet EHR4CR combine et développe plusieurs composants techniques précédemment isolés pour développer une

plateforme pour réutiliser les données des DPI pour soutenir la recherche médicale De Moor et al. [2015].

2.4.1 Plateformes et entrepôts de données cliniques

STRIDE (Stanford Translational Research Integrated Database Environment) STRIDE est un projet mené par l'université de Stanford dont le but est de créer une plateforme standard supportant la recherche clinique translationnelle. Il comporte trois volets : une base de données cliniques, basée sur le standard HL7-RIM, un modèle sémantique basé sur des terminologies et ontologies (SNOMED CT, CIM10 et RxNorm, une terminologie américaine des médicaments disponibles sur le marché nord-américain) et un framework permettant le développement d'applications dédiées à la recherche. Cependant, actuellement aucun projet de déploiement extérieur à l'Université de Stanford n'est prévu et selon les connaissances actuelles, les types de données omiques ne sont pas pris en charge [Iyer et al., 2014; Lowe et al., 2009].

Slim-Prim Slim-Prim (Scientific Laboratory Information Management – Patient-care Research Information Management, Université du Tennessee [Viangteeravat et al., 2009]) est un système permettant de collecter, archiver et distribuer des données de recherche et des données cliniques. Bien que Slim-Prim permette la gestion de données de type micro-arrays, cette plateforme ne gère pas les données de séquences et autres données biomoléculaires.

i2b2 : informatics for integrating biology and the bedside La plateforme Informatics for Integrating Biology and the Bedside (i2b2) a été développée à l'origine au sein du Partners Healthcare System, un vaste système de santé intégré basé à Boston [Murphy et al., 2006]. La première version du code d'i2b2 a été publiée publiquement en 2007 [Kohane, 2011]. Depuis sa création, plus de 200 articles ont été publiés à l'aide de données dérivées des systèmes i2b2. Un objectif majeur de i2b2 est de créer un moyen rentable et efficace d'identifier les patients pour de nombreux types de recherche clinique et translationnelle. i2b2 permet de développer des applications dédiées à la recherche clinique dans le champ génomique grâce à une architecture modulaire [Miyoshi et al., 2013] et donc d'effectuer des recherches portant à la fois sur des données cliniques et des données génomiques [Murphy et al., 2017]. Chaque module est appelé cellule et chaque cellule peut communiquer avec les autres à travers de services Web. Les modules principaux gèrent le stockage des données ou encore la gestion des ontologies. Bien qu'i2b2 soit un outil puissant pour gérer et exploiter des données cliniques pour les fins de la recherche clinique [Johnson et al., 2014], il n'offre pas la possibilité de fouiller les données d'un seul patient.

2.4.2 Plateformes et entrepôts de recherche translationnelle

En sciences translationnelles, plusieurs plateformes sont dédiées au stockage, à la gestion et à l'exploitation de données cliniques et biomoléculaires.

BRISK : Biology-Related Information Storage Kit La plateforme BRISK [Tan et al., 2011] est un assemblage de trois applications Web open source offrant une plateforme cohérente d'intégration et de gestion des données. Il a d'abord été développé pour fournir une solution de partage de données pour les chercheurs du consortium AllerGen (The Allergy, Genes and Environment Network). BRISK peut traiter les informations cliniques du phénotype et de la mutation somatique (polymorphismes à un seul nucléotide). Il fournit aux chercheurs des capacités d'analyse des études d'association (Genome-Wide Association Study (GWAS)) à l'échelle du génome. Cette solution comprend également une application orientée laboratoire qui gère l'échantillon physique, le sujet et les données de conteneur.

caTRIP : Cancer Translational Research Information Platform La plateforme Cancer Translational Research Information Platform (caTRIP) [McConnell et al., 2008] a été développée en tant que composant du projet cancer Biomedical Informatics Grid (caBIG) au début des années 2000 pour permettre aux utilisateurs de faire une requête sur la grille caBIG. caBIG était un programme américain du National Cancer Institute. Son objectif était de développer un réseau open source aux États-Unis pour des échanges sécurisés sur la recherche sur le cancer. Les objectifs de caTRIP sont de permettre aux médecins de trouver des patients avec des profils similaires, d'analyser leurs résultats et de trouver des informations sur les traitements réussis dans la grille de données caBIG. Le système interopère avec plusieurs applications caBIG, y compris (i) le registre des tumeurs, un système clinique utilisé pour collecter des données, (ii) le système d'extraction d'informations textuelles sur le cancer utilisant des terminologies contrôlées, (iii) le caTissue CORE, un dépôt de banque de tissus, (iv) le Cancer Annotation Engine et (v) le caIntegrator, un outil de stockage, d'interrogation et d'analyse de données.

cBio Cancer Genomics Portal Développé au Memorial Sloan-Kettering Cancer Center (MSKCC), le portail cBio Cancer Genomics [Cerami et al., 2012] est une plateforme open source conçue pour faciliter l'accès des chercheurs aux ensembles de données générés par les grands projets de génomique du cancer, comme The Cancer Genome Atlas [21] et le Consortium international du génome du cancer [22]. Il intègre les données cliniques désidentifiées, telles que la description du phénotype, la survie ou les

21. http://cancergenome.nih.gov/
22. http://icgc.org/

intervalles de survie sans maladie, avec des données omiques à haut débit (ADN, ARN messager - mNRA et protéines). De plus, les images de pathologie peuvent être consultées grâce à la visualisation numérique des archives de diapositives[23]. Des fonctionnalités avancées de visualisation, d'analyse et d'exportation sont fournies. La version en ligne publique stocke principalement des ensembles de données à grande échelle sur la génomique du cancer, tandis qu'une instance privée du portail peut être configurée localement par des groupes de recherche disposés à importer leurs propres ensembles de données de recherche.

G-DOC Georgetown Database of Cancer Développé au Lombardi Comprehensive Cancer Center à l'Université de Georgetown, la base de données de Georgetown sur le cancer (G-DOC) [Madhavan et al., 2011] est une plateforme de recherche translationnelle. G-DOC intègre les caractéristiques des patients (par exemple, la démographie, les données de recherche clinique structurée) et les données sur les résultats cliniques avec quatre données omiques majeures à haut débit (ADN, ARNm, microARN et métabolites) dans un environnement unifié. La plateforme contient un large éventail d'outils de bioinformatique et de biologie des systèmes dédiés à l'analyse et à la visualisation des données.

iCOD : Integrated Clinical Omics Database La plateforme iCOD [Shimokawa et al., 2010] a été développée pour combiner des informations cliniques et moléculaires complètes des patients afin de fournir une compréhension holistique des maladies. iCOD peut gérer les données omiques comme les profils d'expression génique et les informations cliniques hétérogènes telles que les phénotypes détaillés, les images de radiologie ou les résultats des tests de laboratoire. Des visualisations intégrées sont fournies pour résumer l'interrelation des données cliniques et omiques et pour représenter des voies plausibles.

iDASH : Integrating Data for Analysis, anonymization and SHaring iDASH [Ohno-Machado et al., 2012] est développé par le National Center for Biomedical Computing. Cette plateforme fournit aux chercheurs américains une infrastructure informatique pour l'intégration des données et l'analyse des données. iDASH distribue également des outils et des algorithmes, axés sur le partage de données de façon sécuritaire.

tranSMART Cette plateforme a d'abord été développée comme une plateforme de collaboration pour les entreprises pharmaceutiques par un consortium privé avant d'être diffusée dans la communauté open source (la Fondation tranSMART est maintenant

23. http://cancer.digitalslidearchive.net/

chargée de la maintenance et du développement) [Szalma et al., 2010]. La plateforme est basée sur le modèle de données open source i2b2. Il est conçu pour aider les scientifiques à développer et à affiner les hypothèses de recherche en étudiant les corrélations entre les données phénotypiques et omiques. tranSMART peut gérer les données structurées à partir d'essais cliniques (démographie, résultats, résultats de laboratoire et phénotypes cliniques) et des données alignées sur les biomarqueurs, tels que les profils d'expression des gènes, les génotypes, les métabolomites et les données protéomiques. Il fournit aux chercheurs des outils d'analyse capables de générer des statistiques descriptives et analytiques avancées.

2.4.3 Comparaison des plateformes existantes

Gestion des données cliniques

BRISK et le portail cBio Cancer Genomics se concentrent principalement sur l'exploration des données omiques. Dans ces plateformes, les données cliniques sont recueillies et stockées pour permettre la catégorisation des échantillons et effectuer des analyses spécifiques (par exemple, type de pathologie pour une analyse de GWAS dans BRISK et des intervalles sans maladie pour une analyse de survie dans le portail cBio). CaTRIP, G-DOC, iCOD, iDASH et tranSMART se concentrent sur l'exploration des données cliniques. iDASH fournit de nombreux outils d'analyse d'image et de Traitement Automatique de la Langue (TAL), et gère les documents en utilisant MIDAS [24], une solution open source. Dans tranSMART, les données phénotypiques sont stockées en utilisant le modèle de données i2b2 constitué d'un schéma en étoile dérivé de paires de valeurs d'entité. G-DOC et iCOD utilisent leur propre format de base de données.

Gestion des données omiques

En ce qui concerne les données omiques, chaque plateforme supporte un ensemble de données spécifiques, en fonction des objectifs initiaux de la plateforme et des besoins des chercheurs qui conduisent le projet. G-DOC prend en charge quatre types de données omiques : ARN messagers, microARN, variation de nombre de copies et spectrométrie de masse de métabolites. En tant que plateforme de recherche translationnelle initialement destinée au domaine du développement de médicaments, tranSMART prend en charge de multiples ensembles de données omiques utiles aux entreprises pharmaceutiques : profils d'expression des gènes, génotypes, profils de protéines sériques, données métabolomiques et protéomiques. La plateforme BRISK est axée sur l'étude d'association GWAS : les polymorphismes à un seul nucléotide sont les seules données omiques supportées. Le portail CBio supporte un large éventail d'ensemble de données omiques produites par des études à grande échelle : données de mutations, modifications du

24. http://midasplatform.org/

nombre de copies, modifications de l'expression de l'ARNm par micro-arrays, valeurs de méthylation de l'ADN et protéines et niveaux de phosphoprotéines. iCOD comprend des données omiques moléculaires telles que l'hybridation génomique comparative et les profils d'expression génique.

Outils de visualisation et d'analyse

Les fonctionnalités d'analyse fournies par le portail CBio Cancer Genomics, G-DOC, iCOD, iDASH et tranSMART s'appuient principalement sur un outil tiers, comme le logiciel statistique R, intégré directement dans les plateformes. Ils fournissent des scripts prêts à l'emploi mettant en œuvre les principaux tests et outils d'analyse utilisés par les chercheurs (y compris - mais non limités - au test t et à l'analyse en composantes principales). De plus, des outils de visualisation multiples sont fournis, via des logiciels tiers (par exemple, Integrative Genome Viewer) ou des développements internes. En plus des outils d'analyse, la plupart des systèmes implémentent des fonctionnalités d'exportation compatibles avec le logiciel SASÕ, R ou Microsoft Excel, permettant une analyse avancée par des experts statisticiens.

Interopérabilité

La plupart des plateformes ne fournissent aucun support pour les terminologies et les ontologies standards. Seuls iDASH et caTRIP ont été conçus pour supporter nativement un ensemble limité de terminologies. tranSMART gère actuellement l'utilisation de terminologies (par exemple la CIM10 ou Logical Observation Identifiers Names and Codes (LOINC) [25]). i2b2 permet d'aligner des terminologies deux à deux à l'aide d'expression régulières. Un environnement collaboratif et sécurisé est également fourni par chaque plateforme, à l'exception de iCOD (information non disponible). Cela permet aux chercheurs de partager et de travailler de manière sélective sur les ensembles de données stockés, ce qui pourrait accélérer le processus de recherche. De manière surprenante, aucune des plateformes ne peut être intégrée dans un cadre global : des formats standard tels que CDISC ODM ou HL7 CDA ne sont pas traités comme format d'entrée et les sorties ne sont pas toujours compatibles avec les pipelines existants d'analyse de bioinformatique. Les développements et déploiements liés à ces plateformes demandent donc des efforts importants.

2.5 L'indexation de textes médicaux

Les données biomédicales, et en particulier les DPI comprennent des informations textuelles, non structurées, comme détaillé dans la partie 2.3.2. Leur exploitation au

25. https://loinc.org

sein d'outils de RI, et plus largement dans des entrepôts de données nécessite un traitement particulier. La plupart des plateformes, comme i2b2, proposent une recherche plein texte pour accéder à l'information contenue dans des textes médicaux. Cependant, lorsque le nombre de documents se multiplie ou que le nombre de termes de la recherche est important, il est avantageux d'utiliser les techniques d'indexation de documents.

L'indexation dans le domaine biomédical implique la réalisation d'un certain nombre de tâches indispensables telles que l'identification de termes médicaux, l'identification d'attributs tels que la négation, l'incertitude, la sévérité, l'identification des relations entre les entités et la mise en correspondance des termes du document aux concepts dans des SOC spécifiques au domaine. Le processus entier dépend d'un certain nombre de processus fondamentaux de TAL tels que la tokenisation et l'analyse syntaxique. Il existe également une forte dépendance vis-à-vis des ressources spécifiques au domaine telles que les dictionnaires médicaux et les SOC tels que l'UMLS.

2.5.1 Concepts, bases et définitions

L'objectif de l'indexation dans le domaine médical est de convertir un document non structuré en information structurée de sorte que l'information puisse ensuite être analysée, agrégée, extraite et traitée de façon automatique.

Un texte médical contient un grand nombre de termes variés : noms de maladies, noms de médicaments, procédures médicales, dispositifs médicaux, résultats de laboratoire, mesures du corps du patient, etc. De plus, chacun de ces termes médicaux ou entités cliniques comporte un certain nombre de modificateurs qui leur sont attachés. Par exemple, une maladie peut être *chronique*, *aiguë*, *légère*, *atypique*, *idiopathique*, etc. De même, un nom de médicament peut être accompagné d'informations supplémentaires telles que la fréquence, l'itinéraire ou la quantité de la dose d'administration.

La forme d'un texte médical n'est pas standardisée. Classiquement, le compte-rendu présente quelques paragraphes de long. Les phrases peuvent être des phrases courtes ou de longues phrases composées. La plupart du texte est narratif sans l'utilisation de constructions stylisées. Par exemple, les phénomènes tels que les doubles négatifs (non inconnus) sont relativement rares. Les documents contiennent des données structurées et non structurées. L'en-tête et le bas du document peuvent contenir des informations sur les patients, sur les médecins et les hôpitaux, sur l'heure et sur la date dans un format structuré. Le corps du document peut également contenir une composante structurée sous la forme d'une liste de diagnostics, d'antécédents, d'allergies connues, etc. Cependant, une grande partie du corps des documents est non structurée. Les médecins décrivent le patient, son état, le diagnostic ou les résultats d'une procédure en texte libre. La plupart des documents sont généralement subdivisés en sections.

Les textes médicaux (comptes-rendus de procédures ou de séjours hospitaliers,

lettres de liaison) posent ainsi de nombreux défis à tout outil d'indexation :

— structure de documents non standard : les textes médicaux n'ont pas de structure standardisée. Ils peuvent être divisés en sections, mais il n'y a pas de standardisation sur le type de sections ou leurs en-têtes ou contenus qui vont varier d'une institution à l'autre, voire d'un médecin à l'autre ;

— technicité du langage : les documents médicaux contiennent un grand nombre de termes médicaux et de vocabulaire spécialisé. Les outils d'indexation généralistes vont ainsi obtenir des résultats peu satisfaisants. Cet aspect concerne par exemple un grand nombre d'actes thérapeutiques (par exemple : la décompression intralabyrinthique par abord des fenêtres, sans laser) ;

— grammaire et syntaxe : souvent, des phrases incomplètes ou des phrases anormalement courtes sont utilisées, semblables à la prise de notes. Par exemple : « Début de fièvre le 26/12 bien tolérée sans point d'appel infectieux. Appétit discrètement altéré » ;

— abréviations : le domaine médical connaît une abondance d'abréviations. Souvent, la même abréviation peut être non médicale ou médicale ou peut s'étendre à des termes différents dans des spécialités médicales différentes selon le contexte et l'intention de l'auteur. Les abréviations sont ainsi difficiles à normaliser, classer ou résoudre ;

— polysémie et synonymie : un seul terme médical peut représenter deux idées différentes basées sur le contexte. C'est ce qu'on appelle la polysémie. Par exemple, *inflammation* peut se référer à un problème de peau, un problème de niveau cellulaire, une activité non médicale, etc. En outre, un concept unique peut être exprimé à travers de nombreux mots différents. C'est ce qu'on appelle la synonymie ;

— erreurs de transcription : la plupart des rapports sont dictés par des médecins et dactylographiés par des tiers. Ceci introduit un large éventail d'erreurs de transcription. Les mots inaudibles sont laissés en blanc. Les homophones tels qu' *antérieur* (avant), *intérieur* (à l'intérieur) créent de la confusion ; des mots épelés de façon similaire sont confondus. Le processus de transcription introduit également un large éventail d'erreurs d'orthographe ou grammaticales.

2.5.2 L'extraction de termes cliniques dans des textes médicaux

L'extraction de termes cliniques est la tâche la plus fondamentale dans l'extraction d'information dans le domaine médical. Elle implique l'extraction de termes médicaux et de phrases à partir de documents. Les termes médicaux peuvent inclure des noms

de maladies, des procédures, des dispositifs médicaux, des noms de médicament, etc. Les termes cliniques peuvent être des unités à un ou plusieurs mots qui se produisent soit de façon contiguë ou non. L'extraction des termes a été largement étudiée et explorée dans la littérature. Un certain nombre d'approches basées sur l'utilisation d'outils statistiques, de règles ou de linguistiques ont été explorées.

Méthodes statistiques

Les méthodes statistiques s'avèrent un choix robuste et généralisable pour de nombreuses tâches de TAL. Elles dépendent fortement de l'obtention de données d'apprentissage. Les modèles de Markov cachés, les systèmes MaxEnt, les champs aléatoires conditionnels sont des modèles courants utilisés pour l'extraction de termes cliniques.

Modèles de Markov cachés Les travaux sur les modèles de Markov cachés tel que ceux de Collier et al. [2000] utilisent la génération d'un modèle de séquence pour la détection des termes cliniques dans le texte. Les probabilités de transition entre types de termes sont utilisées pour prédire la probabilité d'apparition d'un terme suivant son type.

Modèle de Markov de maximum d'entropie Finkel et al. [2005, 2004]; Saha et al. [2009] utilisent une méthode basée sur un modèle de Markov de maximum d'entropie. Cette méthode permet d'utiliser une plus grande variété de fonctionnalités. Les caractéristiques linguistiques telles que la nature grammaticale jouent également un rôle. Les approches basées sur le lexique sont également utilisées pour créer des fonctionnalités supplémentaires.

Champs aléatoires conditionnels Depuis l'introduction des Champ Aléatoire Conditionnel (CAC) [Lafferty et al., 2001], ils ont été un choix populaire pour les tâches d'annotation de textes. Les CAC ont été utilisés pour l'extraction de termes cliniques dans de nombreux travaux [Bodnari et al., 2013; Grouin, 2014; McDonald et Pereira, 2005; Settles, 2004; Tang et al., 2014]. Les caractéristiques utilisées par les modèles de Markov de maximum d'entropie sont également appropriées par les CAC. Un certain nombre de caractéristiques orthographiques telles que la casse, la présence de ponctuation, etc., sont également exploitées pour fournir des indices supplémentaires. Les CAC surmontent le problème du biais des annotations rencontré par les modèles de Markov de maximum d'entropie et ont été théoriquement et empiriquement prouvés être plus robustes et plus précis dans les tâches d'annotation.

Machine à vecteurs de support Les classifieurs à Machine à vecteurs de support (MVS) ont également été utilisés dans des tâches d'extraction de termes cliniques dans

les travaux de Doan et Xu [2010] et Saha et al. [2009]. Par la suite, les MVS ont été modifiées pour des tâches d'annotations sous la forme de classifieurs MVS structurés utilisés dans les travaux de Cogley et al. [2013] et Yamamoto et al. [2003]. Des résultats comparables sont également obtenus avec des classifieurs MVS non structurés.

Combinaison et comparaison des approches Un certain nombre de travaux tentent de combiner plusieurs modèles statistiques entre eux ou avec des approches à base de règles, soit dans un pipeline, soit dans une architecture parallèle combinée pour la majorité. Dehghan [2013] post traite la sortie du CAC pour corriger les erreurs d'identification des limites. Wang et Patrick [2009] combinent les résultats de CAC et de modèles de Markov de maximum d'entropie à l'aide du vote à la majorité. Ils tentent aussi d'utiliser le CAC pour l'identification de termes en post-traitant ensuite les résultats par un système de Markov de maximum d'entropie pour la classification de type de termes. Korkontzelos et al. [2015] utilisent la combinaison d'un modèle de Markov de maximum d'entropie et un perceptron pour agréger les prédictions issues d'une part de dictionnaires et d'autre part de systèmes d'indexation. Des études comparatives telles que celles réalisées par Abacha et Zweigenbaum [2011] révèlent que les systèmes statistiques tels que le CAC obtiennent de meilleurs résultats que les méthodes purement basées sur des règles. De plus, le CAC surpasse également la combinaison d'un système de règle pour l'extraction de termes suivie d'un classifieur MVS pour la classification du type de terme.

Méthodes basées sur des règles

Un certain nombre de méthodes basées sur l'utilisation de règles ont également été proposées pour la reconnaissance de termes cliniques dans des textes médicaux. Les méthodes se répartissent en deux catégories : les approches basées sur la linguistique et les approches basées sur les vocabulaires contrôlés.

Approches linguistiques Les approches basées sur la langue s'appuient généralement sur l'analyse syntaxique. L'analyse syntaxique est réalisée en utilisant un certain nombre de règles supervisées pour identifier les termes. Proux et al. [1998] effectuent un certain nombre d'étapes de filtrage fondées sur des règles pour identifier des entités cliniques. Wilbur et al. [1999] effectuent la segmentation des phrases en utilisant des méthodes basées sur des règles. Ils implémentent également une approche en deux étapes où un système basé sur des règles est suivi d'un classifieur qui identifie le type de terme. De même, Rebholz-Schuhmann et al. [2006] utilisent un certain nombre d'étapes de filtrage utilisant à la fois des principes basés sur des règles et des principes statistiques. Jimeno et al. [2008] incluent un modèle statistique dans le système basé sur des règles en utilisant la fréquence des mots et le nombre de cooccurrences.

Approches basées sur des vocabulaires contrôlés Les approches basées sur les terminologies et ontologies mettent à profit l'utilisation des connaissances médicales avec notamment l'UMLS, la SNOMED CT pour identifier les termes médicaux et caractériser leur type. Par exemple Fan et al. [2013] utilisent l'ontologie SNOMED CT. De même MetaMap [Aronson et Lang, 2010; Mork et al., 2017] est un outil basé sur des règles exploitant le réseau sémantique de l'UMLS. Ces outils sont détaillés dans la section 2.5.4.

Méthodes génériques appliquées au domaine médical

Malgré ses spécificités, la tâche d'extraction de termes cliniques présente une étroite similarité avec la même tâche appliquée dans le domaine général et n'exploitant pas de vocabulaire spécifique. Les points communs [Bikel et al., 1999] incluent la personne, les adresses et les noms d'organisation. Leur identification est une première étape importante dans le domaine de l'extraction de l'information [Nadeau et Sekine, 2007]. La reconnaissance de termes a été un domaine largement étudié et expérimenté. Les approches fondées sur des règles, l'analyse syntaxique, l'utilisation de lexiques fondés sur le Web, l'utilisation d'outils statistiques tels que les CAC [McCallum et Li, 2003; Tkachenko et Simanovsky, 2012], modèles de Markov cachés Zhou et Su [2001] et modèles de Markov de maximum d'entropie [Bender et al., 2003] ont été explorés pour la reconnaissance de termes. Ratinov et Roth [2009] discutent de différents défis de conception pour la reconnaissance de termes en utilisant la représentation d'entités nommées, des schémas d'annotations, de modèles et d'ensembles de caractéristiques, qui se révèlent pertinents pour l'extraction de termes cliniques. Spasić et al. [2015] discutent de l'importance d'une ontologie de domaine pour améliorer les résultats de la reconnaissance de termes lorsqu'il s'agit d'un domaine cible spécifique.

2.5.3 Détection des modificateurs

Les modificateurs sont des termes qui fournissent des informations vitales supplémentaires sur les entités. Un modificateur peut annuler, quantifier ou décrire une entité. Le rôle sémantique d'une entité dans une phrase ne peut être découvert qu'après la combinaison avec ses modificateurs (par exemple : « fièvre augmentée ». Plusieurs approches sont également possibles ici : basées sur l'utilisation de règles ou d'outils statistiques.

Approches basées sur des règles

Un sous-problème important de la détection des modificateurs est le problème de la détection de la négation (« absence de », « pas de »), qui a été largement étudié dans la littérature. Un système de détection de négation basé sur des règles simple,

mais populaire est l'algorithme NegEx proposé par Chapman et al. [2001] qui utilise un lexique de 35 phrases de négation utilisées pour la détection de la négation avec une fenêtre contextuelle de 5 mots pour détecter l'entité négative. Ceci est étendu par Harkema et al. [2009] où, avec des termes déclencheurs, une deuxième liste de termes est utilisée. Cet algorithme a également été exploré pour la langue française, avec de bons résultats [Deléger et Grouin, 2012; Garcelon et al., 2014]. Dans le contexte de la négation, des entités cliniques Patrick et al. [2006] mentionnent une importante classification des négations. Certaines négations sont incluses comme des entités cliniques dans l'UMLS, tandis que d'autres sont des cas de négation classiques dans lesquels la réponse à la négation est disjointe de l'expression de l'entité clinique. Les méthodes basées sur des règles qui dépendent fortement de l'analyse lexicale et linguistique dominent ainsi les efforts de détection de négation. De nombreuses idées peuvent être reprises pour identifier d'autres catégories de modificateurs. Par exemple, une approche d'analyse de syntaxe telle que celle de Gindl et al. [2008] peut être utilisé pour détecter la portée de toute phrase modificatrice.

Approches statistiques

Uzuner et al. [2011] ainsi que de Bruijn et al. [2011] ont proposé une approche basée sur un classifieur MVS et une gamme de fonctions lexicales et syntaxiques pour la classification d'assertions. Clark et al. [2011] divisent la tâche de détection de modificateur en deux étapes : la détection de déclencheurs et la détection du cadre du déclencheur. Ils modèlent chacune de ces étapes en tant que tâches d'annotation en utilisant des CAC.

Approches basées sur l'analyse de dépendance

Sohn et al. [2012] démontrent l'utilité de l'analyse de dépendance pour la détection de négation par une approche basée sur les règles où les motifs créés manuellement sur l'arborescence de l'analyse de dépendance sont utilisés pour identifier les cas de négation. L'utilisation de l'analyse de dépendance avec des classifieurs a été largement implémentée avec la conception de classifieurs spécialisés qui mesurent la similarité des arbres. Les noyaux de convolution [Collins et Duffy, 2001; Moschitti, 2004] sont une telle approche. Sidorov et al. [2012] utilisent des fonctionnalités basées sur l'analyse de dépendance jointes à un classifieur pour la tâche d'idenditification de l'auteur.

2.5.4 Outils d'indexation existants

Application à la langue anglaise

MTI : Medical Text Indexer Développé par la NLM, MTI est un logiciel d'indexation automatique adapté au MeSH. MTI recommande des termes MeSH appropriés à

chaque citation MEDLINE en utilisant le titre et le résumé comme entrée [Aronson et Lang, 2010; Mork et al., 2014]. MTI se compose de deux composants principaux : MetaMap et le module de citations associées à PubMed. MetaMap extrait les concepts biomédicaux du titre et du résumé, puis les aligne aux termes MeSH correspondants, tandis que le module RPC tente de trouver des citations MEDLINE similaires à l'aide d'un algorithme des k-voisins modifié nommé PubMed Related Articles (PRA) [Lin et Wilbur, 2007]. Les termes MeSH de ces citations similaires sont ensuite extraits et combinés avec les termes MeSH par MetaMap. Après quelques étapes de post-traitement, telles que l'application de règles d'indexation, une liste classifiée de termes MeSH est proposée. Depuis 2012, l'usage de MTI par les indexeurs de la NLM ainsi que via son interface Web, MeSH On Demand, a été multiplié par 5 [Mork et al., 2017].

Open Biomedical Annotator Le National Center for Biomedical Ontology (NCBO) développe un système d'accès automatique et basé sur les ontologies aux ressources biomédicales en ligne [Shah et al., 2009b]. Le processus d'indexation est réalisé par le traitement des métadonnées textuelles de diverses ressources telles que les jeux de données de GEO et ArrayExpress pour les annoter et les indexer avec des concepts d'ontologies appropriées. Cette indexation nécessite l'utilisation d'un outil de reconnaissance de concepts pour identifier les concepts d'ontologie dans les métadonnées textuelles de la ressource. Cet outil permet d'obtenir une meilleure précision que MetaMap pour la plupart des ressources et dictionnaires testés, au prix d'un rappel moindre, avec un temps d'exécution amélioré [Shah et al., 2009a].

Système BioASQ Ce système adopte une approche de classification à plat grâce à un classifieur MVS binaire pour chaque annotation dans les données d'apprentissage [Tsoumakas et al., 2013]. Le système utilise un méta modèle (appelé MetaLabeler [Tang et al., 2009]) pour prédire le nombre d'étiquettes d'une instance de test. Pendant la prédiction, tous les classifieurs MVS sont interrogés et les étiquettes sont triées en fonction de la valeur de confiance correspondante. Enfin, le système prédit les n premières étiquettes. Bien que l'approche proposée soit relativement simple, elle nécessite une grande puissance de calcul et un stockage important.

Open Biomedical Annotator Open Biomedical Annotator est basé sur Mgrep [Dai, 2008]. Ce service a accès à un grand dictionnaire de termes biomédicaux issus de l'UMLS et des ontologies NCBO [Musen et al., 2012]. Il s'appuie également sur la structure hiérarchique des ontologies et leurs alignements pour étendre les annotations. Le service est disponible comme un service Web REST pour créer des annotations à partir d'ontologies personnalisées [Jonquet et al., 2009; Shah et al., 2009a].

PoNeDI PoNeDI propose une approche pour l'indexation des documents biomédicaux avec le MeSH et la SNOMED CT qui vise à surmonter la limitation de la correspondance partielle. Cette approche propose de restreindre le processus de racinisation à l'étape du prétraitement. L'étape de l'extraction des descripteurs repose essentiellement sur le modèle possibiliste et combine des méthodes sémantiques et statistiques pour calculer un score estimant la pertinence d'un descripteur dans un document. Les connaissances fournies par l'UMLS sont utilisées pour le filtrage. La méthode de filtrage vise à ne conserver que des descripteurs pertinents. Les expériences menées sur la collection OHSUMED en anglais ont montré de très bons résultats en comparaison avec le baseline (+26.37%) ainsi que les documents de CISMeF en français [Chebil et al., 2015, 2016].

Application à la langue française

FMTI FMTI est un outil d'indexation automatique attribuant des descripteurs MeSH au texte médical en français [Pereira et al., 2008]. Il repose sur une approche multi terminologie impliquant quatre terminologies médicales importantes et les alignements entre elles. Une fois que le texte a été normalisé, les mots vides supprimés, et chaque mot lemmatisé, le groupe de mots obtenu est mis en correspondance indépendamment de l'ordre des mots contre tous les termes MeSH, CIM10, SNOMED CT, Classification Commune des Actes Médicaux (CCAM) et TUV qui ont été traités en de la même façon. Les termes de candidats obtenus sont limités au(x) terme(s) MeSH sémantiquement les plus proches en utilisant des relations interconcept obtenues en alignant les terminologies. En conséquence, la liste finale des termes d'indexation se compose de termes MeSH obtenus directement et les termes MeSH obtenus indirectement à l'aide des relations inter terminologiques.

Peregrine Peregrine Hettne et al. [2010] est un logiciel open source développé par le centre médical de l'Université Erasmus aux Pays-Bas. Il supprime les mots vides et détecte la phrase la plus longue possible pour la faire correspondre à un concept. Il utilise l'outil Lexical Variant Generator de l'UMLS pour réduire l'expression avant la correspondance. Peregrine peut trouver des concepts partiellement superposés, mais il ne peut pas détecter les concepts imbriqués (il ne renvoie que le concept correspondant au terme le plus long). Peregrine n'est pas dédié spécifiquement au français, mais a obtenu les meilleurs résultats lors de la campagne de test 2016 CLEF e-Health dédiée à l'extraction d'information dans des textes cliniques francophones [Névéol et al., 2016; Van Mulligen et al., 2016].

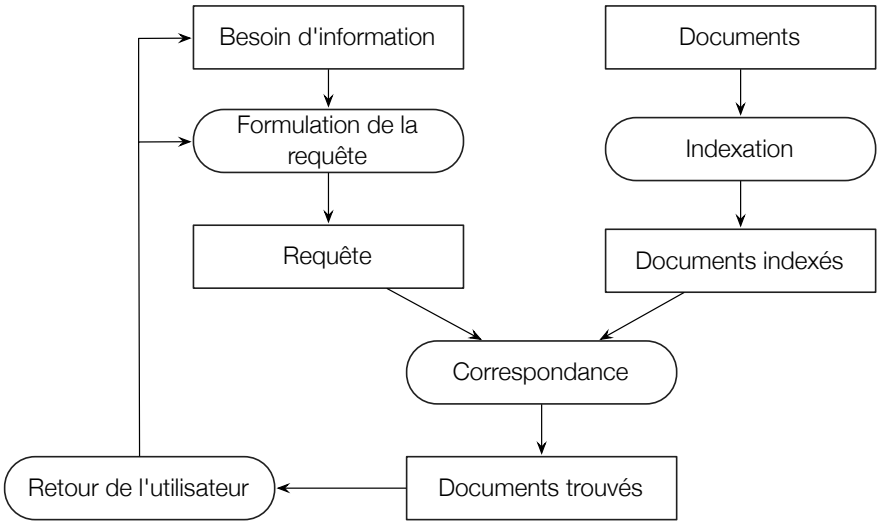

FIGURE 2.8 – Processus général de la recherche documentaire.

2.6 La recherche d'information en santé

En RI, l'index d'un document peut être utilisé dans le processus de recherche. Cette partie décrit les concepts de base de la RI en santé et présente les différents modèles de RI.

2.6.1 Concepts, bases et définitions

La RI est un domaine large et varié, tel que le chercheur en RI Gerard Salton l'indiquait dans sa définition générale originale dans les années 1960 : « Information retrieval is a field concerned with the structure, analysis, organisation, storage, searching and retrieval of information. » [SALTON, 1968].

Cependant, dans ce manuscrit, nous adoptons une conception standard de la RI : un utilisateur ayant un besoin d'information, exprimé à l'aide d'une requête, obtenant une liste triée de documents non structurés, dans l'ordre décroissant d'une mesure de pertinence à sa requête (voir FIGURE 2.8). Les caractéristiques importantes sont ici doubles : les données (documents et requêtes) ne sont pas structurées et il y a une certaine mesure de pertinence (ou d'incertitude) du document à une requête. Cette estimation de la pertinence est naturellement incertaine. Ainsi le domaine de la RI a développé un large corpus de connaissances autour de modèles qui traitent de l'incertitude. Ces modèles peuvent être considérés comme inférentiels de diverses manières : par exemple, l'inférence incertaine qu'un document donné est pertinent pour une description de requête ou pour inférer des termes d'expansion de requête pour augmenter la requête d'origine. Le principe du classement des probabilités [ROBERTSON, 1977] et le principe de l'incertitude logique [VAN RIJSBERGEN, 1986] sont deux exemples qui illustrent l'inférence incertaine centrée sur la RI.

Ensuite, la représentation de l'information en RI est spécifique au contexte. Par représentation, nous entendons à la fois la façon dont l'information est stockée, par exemple dans un index, et comment ces derniers sont utilisés par un modèle de recherche, par exemple en inférant des termes apparentés pour l'expansion de la requête. La RI est spécifique au contexte parce que la représentation est dérivée des données et, par conséquent, reflète étroitement les données spécifiques étant récupérées. Si la représentation est directement dérivée des données, il y a moins de risque d'incompatibilité entre les concepteurs du modèle et les utilisateurs des données. La dérivation de la représentation à partir des données rend également le système relativement léger, plutôt que d'avoir un processus complexe et souvent sujet à l'erreur où les concepteurs construisent manuellement le modèle de domaine. Beaucoup de techniques en RI sont généralement applicables, plutôt que spécifiques au domaine, et peuvent donc être appliquées à n'importe quel domaine. En revanche, l'ontologie peut devoir être adaptée ou ne pas convenir à un domaine autre que celui pour lequel elle a été initialement conçue. Enfin, les modèles RI s'appuient généralement sur des statistiques basées sur les mots et sont donc spécifiquement conçus pour fonctionner avec des données non structurées. Comme les données médicales sont hétérogènes et qu'elles existent en grande partie sous forme de texte libre, les modèles adaptés aux données non structurées sont naturellement applicables. Les approches de RI ont leurs limites. Le principal problème de la recherche sémantique (et en particulier la recherche sémantique des données médicales) est que les modèles RI dépendent des termes comme représentation des documents et des requêtes. L'utilisation d'une représentation à base de termes rend le modèle sensible aux problèmes d'écart sémantique du vocabulaire et de l'inadéquation de granularité. Les modèles RI sont généralement basés sur des statistiques tirées des collections utilisées pour la recherche, généralement aucun recours n'est fait à des sources externes (exception faite de certains modèles RI qui tirent des statistiques supplémentaires de corpus externes [Diaz et Metzler, 2006; Zhu et Carterette, 2012]). L'utilisation de sources externes est très pertinente pour les systèmes de RI médicaux parce que les dossiers médicaux et similaires sont généralement rédigés avec des descriptions de haut niveau qui supposent une connaissance de fond substantielle qui n'est pas explicite.

2.6.2 Historique

Les premiers systèmes de RI apparaissent avec les Sumériens, au début du troisième millénaire av. J.-C., alors qu'ils construisent des zones de stockage et de classification de matériaux manuscrits (inscriptions cunéiformes, un des systèmes d'écriture les plus anciens) pour supporter le travail de groupes sociaux variés [Valentine, 2012]. L'avènement du papier et de la presse imprimée a amené le développement de systèmes pour stocker, gérer et récupérer l'information. En 1945, Vannevar Bush critique le caractère artificiel des systèmes d'indexation contemporains. Il théorise un système fonctionnant

par association à la manière de l'esprit humain et mécanisé afin qu'il puisse être utilisé avec rapidité et souplesse [Bush, 1945]. C'est la première conception d'un système de RI automatisé. Le terme de RI fut inventé un peu plus tard, en 1950, par Calvin Mooers [Mooers, 1950]. Depuis la fin des années 1950, le domaine de la RI a évolué au rythme de nombreux travaux : les travaux de Luhn [Luhn, 1957], le système SMART créé par G. Salton et ses étudiants dans le cadre desquels ont été développés des concepts importants comme le modèle d'espace vectoriel et la réinjection de pertinence [Salton, 1971], le modèle Cranfield [Cleverdon, 1967], le développement de l'Inverse Document Frequency (IDF) [Jones, 1972] et les modèles probabilistes par Robertson [Robertson et Jones, 1976] et Croft [Croft et Harper, 1979; Turtle et Croft, 1991]. En 1992 s'initie la conférence Text REtrieval (TREC) qui fournit l'infrastructure nécessaire à une évaluation à grande échelle et permit l'amélioration de l'existant et l'émergence de nouveaux modèles. Dans les années 1990, avec le développement du Web, le domaine de la RI s'est élargi. L'intérêt accru pour le grand public ainsi que l'augmentation de l'information disponible et des besoins des utilisateurs ont contribué à l'émergence de nouveaux domaines hors du champ de la simple recherche documentaire parmi lesquels la recherche de questions, la recherche multilingue, la détection et le suivi des sujets, la synthèse, la recherche multimédia.

Malgré les développements récents et nombreux, la RI est loin d'être un « problème résolu » [Callan et al., 2007]. D'un côté, l'information est produite à un volume plus important que jamais et de plus, les façons dont les gens produisent, recherchent, gèrent et utilisent l'information évoluent rapidement.

2.6.3 Modèles de recherche d'information

Ayant donné une définition générale de la RI, dont ses avantages et limitations, nous allons maintenant considérer les modèles de RI. Les modèles qui ont été développés pour la tâche de RI sont le modèle booléen et des modèles statistiques incluant le modèle d'espace vectoriel et le modèle probabiliste.

Le modèle booléen

Le modèle booléen est le modèle historique en RI. Ce modèle est basé sur la logique booléenne et la théorie des ensembles : à la fois les documents à rechercher et la requête de l'utilisateur sont conçus à partir d'un même ensemble de termes. La recherche est donc basée sur les documents contenant ou non les termes de la requête. Par exemple, la requête contenant l'unique terme *santé* définit l'ensemble de documents indexés avec le terme *santé*. Les termes d'une requête peuvent être mis en correspondance avec les ensembles de documents contenant ces termes et combinés pour créer de nouveaux ensembles de documents en utilisant les opérateurs logiques ; le produit logique `AND`,

la somme logique `OR` et la différence logique `NOT`. Une requête combinant deux termes à l'aide de l'opérateur `AND` produira un ensemble de documents inférieur ou égal aux ensembles de documents de chacun des termes de la requête, c'est-à-dire leur intersection. Une requête combinant deux termes avec l'opérateur `OR` produira un ensemble de documents supérieur ou égal aux ensembles de documents de chacun des termes de la requête, c'est-à-dire leur union.

Ce modèle a plusieurs avantages expliquant son usage répandu. Il est aisé à implémenter et efficace [Frakes et Baeza-Yates, 1992]. De ce fait, c'est le modèle standard utilisé à grande échelle par les systèmes de RI opérationnels ainsi que de nombreux services en ligne. Il permet également aux utilisateurs d'exprimer des contraintes structurelles et conceptuelles pour décrire des caractéristiques linguistiques (des synonymes avec l'opérateur `OR` par exemple) [Marcus, 1991]. La logique booléenne permet une expression suffisante et claire et s'avère particulièrement efficace si la requête nécessite une sélection exhaustive. De plus, l'approche booléenne permet d'appliquer des techniques de modifications de la requête (expansion ou limitation). Enfin, le modèle booléen est particulièrement efficace dans les phases finales de la RI, car les relations et concepts peuvent être représentés de façon exacte.

Il présente également plusieurs limites, intrinsèquement liées à la nature même des opérateurs booléens. Les utilisateurs peuvent avoir des difficultés à formuler une requête en utilisant des opérateurs booléens pour plusieurs raisons. D'abord, la différence conceptuelle entre la définition booléenne des termes *and, or, not* et leurs significations en langage naturel. L'opération `A AND B` sera alors comprise et utilisée comme une somme entre les ensembles de documents correspondants, plutôt que leur intersection. De la même façon, le `OR` logique sera confondu avec un *ou* exclusif, plus proche de son sens dans la langue parlée. Ainsi, des erreurs dans la construction des requêtes seront liées à ce décalage. De plus, la construction de requête nécessite à l'utilisateur d'être familier avec les concepts de priorités, et de groupes définis à l'aide de parenthèses. Enfin, l'utilisateur doit également gérer les nombreuses façons de structurer une requête, du fait des possibilités combinatoires augmentant avec le nombre de concepts utilisés [Lancaster, 1993]. Ensuite, seuls les documents satisfaisant une requête de façon exacte sont récupérés. L'opérateur `AND` ne distinguera pas le cas où aucun terme n'est retrouvé du cas où seul un des termes est retrouvé. Ainsi peu de documents seront retrouvés lorsque la requête combine plus de trois ou quatre termes. De la même façon, l'opérateur `OR` aura souvent l'inconvénient de produire un trop grand nombre de résultats. Il est ainsi difficile de contrôler le nombre de documents retrouvés et les utilisateurs seront souvent confrontés à un résultat de recherche soit négatif, ou un nombre de résultats trop important. Enfin, le modèle booléen ne fournit aucun système de classement par pertinence des résultats retrouvés [Belkin et Croft, 1992].

Le modèle booléen étendu

Plusieurs méthodes ont été développées pour étendre le modèle booléen afin de résoudre les problèmes mentionnés notamment l'introduction d'une pondération des termes de la requête ainsi qu'un système de classement des résultats [Fox et al., 1992; Fox et Sharan, 1986; Sachs, 1976; Salton et al., 1983].

La méthode P-norm développée par Salton et al. [1983] permet d'associer aux termes de la requête et des documents des poids calculés en utilisant des statistiques de fréquence de terme avec les procédures de normalisation appropriées. Ces poids normalisés peuvent être utilisés pour classer les documents dans l'ordre de la distance décroissante par rapport au point $(0, 0, ..., 0)$ pour une requête `OR`, et dans l'ordre de la distance croissante du point $(1, 1, ..., 1)$ pour une requête `AND`. De plus, les opérateurs booléens se voient associer un coefficient P pour indiquer le degré de rigueur de l'opérateur (de 1 pour le moins strict à l'infini pour le plus strict, c'est-à-dire le cas booléen). La norme P utilise une mesure basée sur la distance et le coefficient P détermine le degré d'exponentiation à utiliser. L'exponentiation est un calcul coûteux, en particulier pour les valeurs P supérieures à un.

Dans la théorie de la correspondance approximative (fuzzy match), un élément a un degré variable d'appartenance à un ensemble au lieu du choix d'appartenance binaire traditionnel. Le poids d'un terme d'index pour un document donné reflète le degré auquel ce terme décrit le contenu d'un document. Par conséquent, ce poids reflète le degré d'appartenance du document dans l'ensemble approximatif associé au terme en question. Le degré d'appartenance à l'union et l'intersection de deux ensembles approximatifs est égal au maximum et au minimum, respectivement, des degrés d'appartenance des éléments des deux ensembles. Dans le modèle « Mixed Min and Max » développé par Fox et Sharan [1986], les opérateurs booléens sont adoucis en considérant la similitude de la requête-document comme une combinaison linéaire des poids minimum et maximum des documents.

Le modèle vectoriel

Le modèle vectoriel est un modèle algébrique pour représenter des documents textuels (ou d'autres objets d'une manière générale) comme des vecteurs d'identifiants, et dans notre cas d'index de termes. Il est utilisé aussi bien dans la RI, dans l'indexation ainsi que dans les systèmes de classements. Il apparaît pour la première fois dans le système de RI SMART dans les années 1960 [Salton, 1971].

Les documents et les requêtes sont représentés comme des vecteurs dans un espace euclidien à plusieurs dimensions. Chaque dimension correspond à un terme distinct (voir Figure 2.9).. Si un terme apparaît dans le document, sa valeur associée dans le vecteur est non nulle. Différentes manières de traiter ces valeurs existent, aussi connues sous le terme de poids. L'une des méthodes les plus connues est la méthode Term

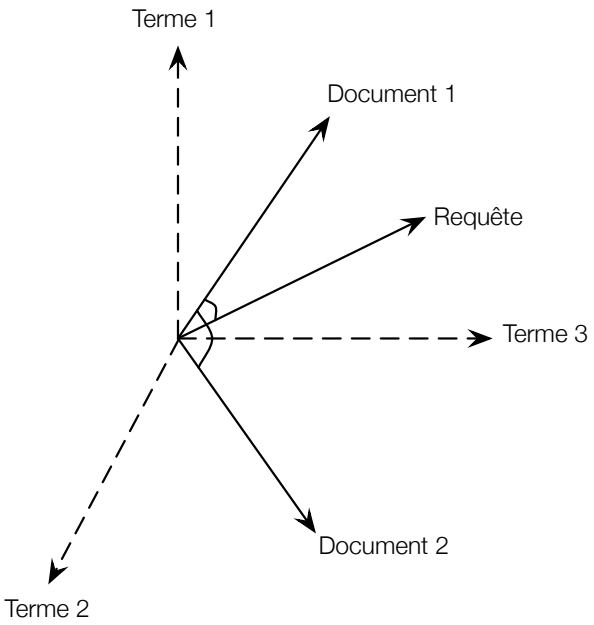

FIGURE 2.9 – Représentation vectorielle de l'espace de documents.

Frequency-Inverse Document Frequency (TF-IDF).

La définition du terme dépend de son application. Les termes sont généralement des mots, mot-clés ou des groupes nominaux. Si l'on choisit des mots comme termes, la dimension du vecteur est le nombre de mots contenus dans le vocabulaire, c'est-à-dire, le nombre de mots distincts du corpus. Les opérations sur les vecteurs peuvent être utilisées pour comparer les documents avec des requêtes.

Si la représentation de l'index d'un document est un vecteur $\vec{d} = (d_1, d_2, \ldots, d_m)$ dans lequel chaque terme d_k $(1 \leq k \leq m)$ est associé à un terme index, et si la requête est un vecteur similaire $\vec{q} = (q_1, q_2, \ldots, q_m)$ dans lequel les composants sont associés avec les mêmes termes, alors leur degré de correspondance est donné par leur similarité. Classiquement on pourra utiliser le simple produit scalaire entre les deux vecteurs, ou encore la mesure de l'angle qu'ils forment, donnée par la relation (2.1) :

$$\text{score}(\vec{d}, \vec{q}) = \frac{\sum_{k=1}^{m} d_k \times q_k}{\sqrt{\sum_{k=1}^{m}(d_k)^2 \times \sum_{k=1}^{m}(q_k)^2}} \tag{2.1}$$

$$\text{score}(\vec{d}, \vec{q}) = \frac{\sum_{k=1}^{m} d_k \times q_k}{\frac{1}{2}\left(\sum_{k=1}^{m}(d_k)^2 + \sum_{k=1}^{m}(q_k)^2\right)} \tag{2.2}$$

$$\text{score}(\vec{d}, \vec{q}) = \frac{\sum_{k=1}^{m} d_k \times q_k}{\sum_{k=1}^{m}(d_k)^2 + \sum_{k=1}^{m}(q_k)^2 - \sum_{k=1}^{m} d_k \times q_k} \tag{2.3}$$

$$\text{score}(\vec{d}, \vec{q}) = \frac{\sum_{k=1}^{m} d_k \times q_k}{min\left(\sum_{k=1}^{m} d_k, \sum_{k=1}^{m} q_k\right)} \tag{2.4}$$

D'autres mesures existent telles que le coefficient de Dice (2.2), la mesure de Jaccard

(2.3), ainsi que la mesure de recouvrement (2.4).

La principale limite du modèle vectoriel est qu'il ne définit pas les valeurs des composants du vecteur. Le problème de l'assignation de valeurs appropriées aux composants du vecteur est appelé pondération des termes. Les premières expérimentations de Salton Salton [1971] et Salton et Yang [1973] ont montré que la pondération des termes n'est pas un problème trivial. Ils ont suggéré une mesure appelée TF-IDF, une combinaison de la fréquence du terme TF, c'est-à-dire le nombre d'occurrences d'un terme dans un document, et IDF, la fréquence inverse de document, c'est-à-dire, le nombre de documents qui contiennent le terme. De nombreux algorithmes modernes sont des versions de cette méthode, la méthode originale de Salton obtenant des résultats contrastés, dans certains cas, moins bons qu'une pondération par la seule fréquence inverse de document. Moins le terme apparaît dans un document, plus le terme est discriminant entre les documents et, par conséquent, plus il est utile dans la RI. La fréquence du document inversé IDF est calculée par la relation (2.5)

$$idf_i = \log \frac{N}{n_i} \tag{2.5}$$

Où idf_i est la fréquence du document inverse pour le terme i, N le nombre total de documents dans la collection et n_i le nombre de documents qui contiennent le terme i.

Le modèle probabiliste

Le modèle probabiliste est basé sur le principe de classement probabiliste. Bien que l'idée d'introduire un classement par les probabilités de pertinence ait été formulée initialement par Maron et Kuhns, c'est Stephen Robertson qui établit le principe de classement probabiliste en 1977 [Robertson, 1977] :

> Un système qui retourne pour chaque requête une liste de documents dans l'ordre décroissant de la probabilité que le document soit utile à l'utilisateur ayant soumis la requête, en supposant que ces probabilités soient estimées aussi exactement que possible à partir de toute l'information disponible, aura la meilleure performance possible sur la base de cette information.

Le principe tient ainsi compte du fait qu'il existe une incertitude dans la représentation des besoins d'information et des documents. Il peut y avoir une variété de sources de preuves qui sont utilisées par les méthodes probabilistes, et la plus courante est la répartition statistique des termes dans les documents pertinents et non pertinents.

Définissons une expérience pour laquelle nous prenons un document de la collection au hasard. Si nous connaissons le nombre de documents pertinents dans la collection, par exemple 100 documents sont pertinents et nous connaissons le nombre total de documents dans la collection, par exemple 1, alors ce quotient définit la probabilité de pertinence $P(R = 1) = 100/1000000 = 0,0001$. Supposons en outre que $P(D_k)$ est la

probabilité qu'un document contienne le terme k avec l'espace échantillon $[0, 1]$, ($0 =$ le document ne contient pas de terme k, $1 =$ le document contient le terme k), puis nous utiliserons $P(R, D_k)$ pour désigner la distribution de probabilité conjointe avec les résultats (0, 0), (0, 1), (1, 0) et (1, 1) et nous utiliserons $P(R|D_k)$ pour désigner la distribution de probabilité conditionnelle avec les résultats 0, 1. Ainsi, $P(R = 1|D_k = 1)$ est la probabilité de pertinence si l'on considère les documents qui contiennent le terme k.

Approche classique Stephen Robertson et Karen Spärck-Jones ont fondé leur approche sur ce principe [Robertson et Jones, 1976]. Ils ont suggéré de classer les documents par $P(R|D)$, c'est-à-dire la probabilité de pertinence R étant donnée la description de contenu du document D. Notons que D est ici un vecteur de composantes binaires, chaque composante représentant typiquement un terme. Dans le modèle probabiliste, la probabilité $P(R|D)$ doit être interprétée comme suit : il peut y avoir plusieurs, par exemple 10, documents qui sont représentés par le même D. Si 9 d'entre eux sont pertinents, alors $P(R|D) = 0,9$. En pratique, la règle de Bayes (2.6) sur les probabilités de probabilité $P(R|D)/P(\overline{R}|D)$ est utilisée, où $\overline{R}$ désigne la non-pertinence. Les probabilités permettent d'ignorer $P(D)$ dans le calcul tout en fournissant un classement par la probabilité de pertinence. En outre, l'indépendance entre les termes est supposée.

$$\frac{P(R|D)}{P(\overline{R}|D)} = \frac{P(D|R)P(R)}{P(D|\overline{R})P(\overline{R})} = \frac{\prod_k P(D_k|R)P(R)}{\prod_k P(D_k|\overline{R})P(\overline{R})} \tag{2.6}$$

Maron et Kuhns [1960] suggèrent en 1960 que la probabilité $P(R)$ pourrait être définie par les statistiques d'usage d'un document. Cette idée est à l'origine du classement par popularité, très utilisée aujourd'hui sur le Web [Joachims et al., 2005].

L'exploitation complète de cette approche nécessite deux éléments : des exemples de documents pertinents et de longues requêtes. Des documents pertinents sont nécessaires pour calculer $P(Dk|R)$, c'est-à-dire la probabilité que le document contienne le terme k étant donné la pertinence. Des requêtes longues sont nécessaires, car cette approche ne distingue que présence et absence d'un terme dans les documents, et ainsi, le nombre de valeurs distinctes pour un document est restreint pour des requêtes courtes.

L'une des applications de cette approche est la méthode BM25, mise en pratique initialement dans le système de RI Okapi dans les années 90 [Robertson et Walker, 1994]. La méthode BM25 utilise la fréquence du terme dans le document, la fréquence du terme dans le corpus et la longueur du document pour estimer la pertinence. Étant donné une requête Q, contenant les termes $q_1, \ldots, q_n$, la fonction de classement d'un document D est donnée dans l'équation (2.7).

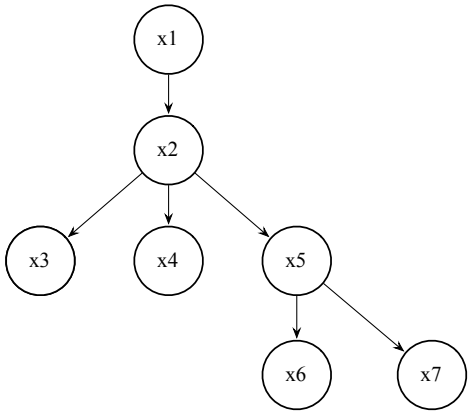

FIGURE 2.10 – Un arbre de dépendance entre termes.

$$RSV(D,Q) = \sum_{q \in Q} \frac{tf_{q,D}\,(k_1+1)}{tf_{q,D} + k_1\left(1 - b + b\frac{|D|}{|D_{avg}|}\right)} \log \frac{|C| - d_{f_q} + 0,5}{d_{fq} + 0,5} \tag{2.7}$$

La fraction à gauche est la composante de pondération de termes, où $tf_{q,D}$ est la fréquence du terme t dans le document D, $|D|$ est la longueur du document et $|D_{avg}|$ est la longueur moyenne du document. La fraction de droite correspond à l'IDF, où $|C|$ est le nombre de documents dans la collection et d_{f_q} est le nombre de documents contenant le terme q. BM25 a deux paramètres libres, b et k_1, qui contrôlent respectivement l'effet de la fréquence des termes et de la longueur du document. Cet algorithme de pondération est aujourd'hui toujours couramment utilisé [SPARCK JONES et al., 2000].

Réseaux bayésiens Turtle et Croft [TURTLE et CROFT, 1989, 1991] ont introduit en RI l'utilisation de réseaux bayésiens [JENSEN et JENSEN, 2001]. Conceptuellement, les réseaux bayésiens reposent sur des graphes orientés pour indiquer les dépendances probabilistes entre les variables (voir FIGURE 2.10), et ont conduit au développement d'algorithmes sophistiqués. Turtle et Croft ont utilisé un réseau pour mieux modeler les dépendances complexes entre un document et les besoins d'information d'un utilisateur.

Le modèle se décompose en deux parties : un réseau représentant la collection de documents et un réseau de requêtes. Le réseau de documents est vaste, mais peut être prétraité : il lie les documents aux termes et concepts. Les concepts sont une extension des termes apparaissant dans le document. Le réseau de requêtes est relativement petit, mais un nouveau réseau doit être construit chaque fois qu'une requête arrive, puis relié au réseau de documents. Le réseau de requêtes fait correspondre les termes de la requête à des sous-expressions (construites à l'aide de versions probabilistes des opérateurs `AND` et `OR`) correspondant aux besoins d'information de l'utilisateur.

Le résultat est un réseau probabiliste flexible qui peut généraliser divers modèles booléens et probabilistes plus simples. Le système a permis une RI efficace à grande échelle, et a été la base du système de recherche de texte InQuery, construit à l'Uni-

versité du Massachusetts [Callan et al., 1992]. Ce système s'est très bien comporté dans les évaluations de TREC et a été vendu commercialement pendant un certain temps. Cependant, le modèle utilisait encore diverses approximations et hypothèses d'indépendance pour rendre possible l'estimation des paramètres et le calcul [Callan et al., 1995].

Modèles de langue Les modèles de langue ont été appliqués à la RI par un certain nombre de chercheurs à la fin des années 1990 [Hiemstra et Kraaij, 1998; Ponte et Croft, 1998]. Ils proviennent de modèles probabilistes de langue développés pour les systèmes automatiques de reconnaissance vocale au début des années 80.

Pour chaque document d'une collection, un modèle statistique en deux étapes définit la probabilité de générer la demande de l'utilisateur. Les documents sont classés selon cette probabilité. Si une requête est saisie, le système utilise d'abord le modèle de formulation de requêtes pour déterminer pour chaque mot de la requête les termes qui ont pu le générer. Il en résulte une requête structurée qui représente toutes les requêtes qui ont généré la demande. Dans la deuxième étape, le système utilise le modèle de correspondance de chaque document pour calculer la probabilité que le document ait généré l'une des requêtes représentées par la requête structurée. Les deux parties ont des objectifs qui sont similaires aux deux parties des modèles de reconnaissance vocale. Les deux modélisent le signal observé, respectivement la demande de l'utilisateur et l'onde sonore. La requête structurée qui représente toutes les requêtes qui ont pu générer la demande peut être comparée à un réseau de mots dans la reconnaissance vocale [Ephraim et Rabiner, 1990].

Dans le modèle de vraisemblance de la requête, un modèle de langage séparé est associé à chaque document d'une collection. Les documents sont classés en fonction de la probabilité de la requête Q dans le modèle de langue du document $P(Q|M_d)$. Généralement, le modèle Unigram est utilisé à cette fin.

Le manque de données est un problème majeur dans la construction des modèles de langue. La plupart des séquences de mots possibles ne seront pas observées lors de la phase d'apprentissage. Une solution consiste à faire l'hypothèse que la probabilité d'un mot dépend uniquement des n mots précédents, hypothèse appliquée dans les modèles n-gram.

Le modèle possibiliste Le modèle possibiliste est une théorie de l'incertitude dédiée au traitement des informations incomplètes. Dans une large mesure, il est comparable à la théorie des probabilités. Il diffère de ce dernier par l'utilisation d'une paire de fonctions, les mesures de possibilité et de nécessité, au lieu d'une seule. En outre, il n'est pas additif et a un sens sur les structures ordinales. Le nom « Theory of Possibility » a été inventé par Zadeh et al. [1978], qui a été inspiré par un article de Gaines et Kohout [1975]. De l'avis de Zadeh, les distributions de possibilités étaient censées

fournir une sémantique graduelle aux énoncés en langage naturel. Il peut être considéré soit comme une version non numérique de la théorie des probabilités, soit comme un cadre de raisonnement avec des probabilités extrêmes, soit comme une simple approche du raisonnement avec des probabilités imprécises [Dubois et Prade, 1998].

2.6.4 La reformulation de requête

Souvent, les termes de la requête sont liés à des termes utilisés pour indexer des documents, mais qui ne font pas partie de l'index du document. Cela motive le développement de techniques d'expansion des requêtes. Ces méthodes sont utilisées pour améliorer la précision des résultats de recherche avec des termes de requête alternatifs ou supplémentaires (synonymes ou autres relations sémantiques). Cette technique est très utilisée, probablement en raison du grand nombre de représentations de concepts de santé disponibles.

Les premiers travaux importants dans ce domaine ont été effectués par Voorhees [1994], qui a étudié si la RI pouvait être améliorée en élargissant les requêtes avec les synonymes de WordNet. Les résultats ont montré qu'il est très difficile de sélectionner automatiquement des termes d'expansion appropriés. Des termes, choisis manuellement, ont cependant réussi à améliorer les résultats. Les résultats de Voorhees ont également montré que la performance en RI basée sur le concept dépend fortement du modèle de domaine spécifique ou de l'ontologie utilisée. Les applications générales (celles qui utilisent WordNet ou Open Directory) ont du mal à surpasser les systèmes basés sur des mots-clés [Egozi et al., 2011; Ravindran et Gauch, 2004; Voorhees, 1994]. En conséquence, l'expansion de requête s'est initialement peu développée. Cependant, les applications biomédicales (qui utilisent des ontologies spécifiques à ce domaine) se sont constamment améliorées [Koopman et al., 2011; Liu et Chu, 2007; Zhou et al., 2007]. Après la méthode d'expansion des requêtes de Voorhees, les modèles suivants ont tenté d'améliorer le modèle de requête avec des représentations basées sur des concepts. Cela a été fait dans le but de résoudre le décalage sémantique. Les termes de requête sont normalisés à des concepts, la motivation étant qu'un concept encapsule toutes les variantes lexicales d'un même terme en une seule entité. Au moment de la recherche, peu importe la variante de terme utilisée, car chaque variante du terme correspondra au même concept global. Zhong et Huang [2006] ont appliqué avec succès cette approche à la recherche de données génomiques, bien qu'ils aient limité les concepts pour représenter des variantes lexicales de noms de gènes dans les données dans le volet Genomics du challenge TREC. Sur la base de ce travail initial, des tentatives ultérieures ont été faites pour utiliser des concepts dans des modèles de langage probabilistes. Trieschnigg [2010]; Trieschnigg et al. [2010] ont construit un modèle de langage de requêtes comme une distribution de probabilité sur des concepts. Ces approches ont démontré des améliorations statistiquement significatives dans la RI, mais avec

des gains limités et souvent une approche très spécifique à la tâche à accomplir (par exemple, seulement applicable à la recherche de données génomiques). La littérature fait référence à un facteur de réussite critique, à savoir les approches qui combinent les statistiques basées sur le corpus et les connaissances du domaine. Ce fut la découverte de Stokes et al. [2009], qui a mené une enquête approfondie sur les critères nécessaires à une expansion réussie des requêtes. Bien que spécifique au domaine de la génomique, un certain nombre de leurs résultats peuvent être généralisés à la RI médicale.

Les méthodes qui combinent les statistiques basées sur le corpus et les connaissances du domaine ont été les plus efficaces. Un certain nombre de pistes ont été explorées exploitant davantage de méthodes axées sur les données dans le cadre d'une approche fondée sur des concepts. Les concepts peuvent être intégrés dans des modèles de langage probabilistes pour créer une représentation basée sur le concept de la requête. Il s'agit d'une pré récupération effectuée et donc indépendante du contenu du document récupéré. Meij et al. [2010]; Trieschnigg [2010] ont étendu ce travail en utilisant le retour de l'utilisateur pour générer un modèle de requête basé sur des concepts mis à jour. Leurs résultats ont montré que l'intégration des statistiques basées sur le corpus avec la connaissance du domaine était la composante clé pour l'expansion réussie des requêtes.

Liu et Chu [2007] ont constaté que les requêtes médicales pouvaient être adaptées à un certain nombre de scénarios différents, par exemple les traitements, le diagnostic, les symptômes. L'UMLS fournit les connaissances de domaine pertinentes sur ces scénarios globaux. Les méthodes standard d'extension de requêtes statistiques peuvent être appliquées, mais filtrées sur la base de concepts correspondant à ces scénarios médicaux spécifiques. Cette heuristique combinée statistique et ontologique a surpassé à la fois une approche statique pure et une approche purement ontologique. Zhou et al. [2007] a franchi une étape supplémentaire en intégrant des types sémantiques. En utilisant des concepts, des types sémantiques et des statistiques de corpus, ils ont pu dériver les relations implicites entre les concepts, et les utiliser pour l'expansion des requêtes. Cette approche s'est révélée la plus efficace lors du volet Genomics du challenge TREC 2006 [Zhou et al., 2006]. Les travaux décrits sont principalement centrés sur le domaine de la génomique, qui est un scénario de recherche très précis. Les requêtes sont fournies sous la forme « Gène (1..n) Processus biologique (1..m) » et la tâche est de retourner des informations pertinentes sur les gènes spécifiques. Par conséquent, un certain nombre de méthodes sont spécifiques à ce domaine et ne peuvent pas être appliquées à d'autres scénarios. Cependant, ils soulignent que les approches couronnées de succès utilisent généralement à la fois les connaissances du domaine et les méthodes statistiques, basées sur les données.

2.7 Évaluation des systèmes de recherche d'information en santé

L'objectif principal de l'évaluation des Système de Recherche d'Information (SRI) est de mesurer la façon dont un besoin d'information d'un utilisateur est satisfait par une liste de documents renvoyés pour une requête spécifique. Il existe une longue histoire d'évaluation empirique en RI et des évaluations robustes des SRI sont enracinées dans la communauté RI [Cleverdon, 1991]. Cette partie passe en revue certains des travaux connexes dans l'évaluation.

2.7.1 Métriques

L'évaluation en RI repose sur des mesures statistiques de l'efficacité du système de RI. La plupart des mesures sont conçues pour quantifier deux éléments : la précision et le rappel [Manning et al., 2008]. La précision est la mesure du nombre de documents pertinents récupérés par le SRI parmi tous les documents, ou plus formellement (équation 2.8) :

$$\text{precision}_{\text{SRI}} = \frac{|D_{rel} \cap D_{ret}|}{|D_{ret}|} \tag{2.8}$$

où D_{ret} représente l'ensemble des documents récupérés par les SRI et D_{rel} l'ensemble des documents pertinents.

Parallèlement, le rappel est la mesure du nombre de documents pertinents récupérés parmi tous les documents pertinents (équation 2.9) :

$$\text{rappel}_{\text{SRI}} = \frac{|D_{rel} \cap D_{ret}|}{|D_{rel}|} \tag{2.9}$$

En RI médicale, il existe différents cas d'utilisation nécessitant soit la maximisation de la précision, soit du rappel. Un scénario commun où les deux sont nécessaires est le cas de la recherche de patients éligibles à l'inclusion dans les essais cliniques [Voorhees et Hersh, 2012]. Les essais cliniques sont conduits pour le développement de nouveaux médicaments ou procédures médicales. Trouver des patients satisfaisants aux critères pour mener un essai clinique peut être vu essentiellement comme un problème de RI - les critères d'inclusion d'essai clinique étant le besoin d'information et les dossiers des patients étant le corpus de documents. Pour un besoin d'information concernant la recherche de patients atteints d'une maladie rare, il est très important que le SRI renvoie tous les documents pertinents (rappel maximal). Dans ce cas, l'utilisateur préférerait inclure de nombreux patients non pertinents que manquer l'un des rares patients pertinents dans son groupe. Inversement, pour une maladie commune, où il ya un grand nombre de patients pertinents, la précision est importante. Les utilisa-

teurs du SRI n'ont pas besoin de tous les documents pertinents, mais ils ne souhaitent pas lire des documents non pertinents. La précision et le rappel sont incorporés dans un certain nombre de mesures d'évaluation standard.

La précision à certaines positions de classement - par exemple la précision à 10 - mesure le nombre de documents pertinents jusqu'à la position de rang stipulée. Étant donnée une position de rang n, la précision @ n est (équation 2.10) :

$$\text{precision @ } n = \frac{\sum_{i=1}^{n} \text{rel}(d_i)}{n} \tag{2.10}$$

où $\text{rel}(d_i)$ représente une fonction, telle que $\text{rel}(d_i) = 1$ si le document d_i est pertinent et $\text{rel}(d_i) = 0$ sinon.

Cette mesure serait la plus appropriée lorsque la maximisation de la précision est importante, par exemple dans le cas de la recherche de maladies ou de maladies courantes. Le rappel peut également être mesuré à des positions de classement spécifiques (équation 2.11) :

$$\text{rappel @ } n = \frac{\sum_{i=1}^{n} \text{rel}(d_i)}{R} \tag{2.11}$$

où R représente le nombre total de documents pertinents.

La position de rang n est souvent définie sur le nombre total de documents renvoyés. Plutôt que de disposer de deux mesures distinctes pour la précision et le rappel, il est intéressant d'avoir une mesure qui englobe les deux composantes. La précision moyenne (AP) permet ceci et est calculée comme suit (équation 2.12) :

$$AP = \frac{1}{R} \sum_{n=1}^{N} P@n \tag{2.12}$$

où R représente le nombre de documents pertinents et N le nombre de documents renvoyés.

On peut également calculer la précision moyenne à travers l'ensemble des requêtes Q, mesure appelée Mean Average Precision (MAP) :

$$MAP = \frac{\sum_{q \in Q} AP(q)}{Q|} \tag{2.13}$$

La MAP est une mesure largement utilisée dans l'évaluation de la RI (équation 2.13). Cependant, elle s'appuie sur l'hypothèse de l'exhaustivité : tous les documents pertinents au sein d'une collection d'essais ont été identifiés [Cleverdon, 1991]. Lorsque cette hypothèse n'est pas vérifiée (c'est-à-dire qu'un nombre important de documents pertinents ne sont pas évalués), les mesures d'évaluation standard décrites ci-dessus ne sont pas robustes. Pour faire face à cette situation, Buckley et Voorhees [2004] ont présenté la mesure d'évaluation bpref qui est conçue pour traiter des informa-

tions incomplètes (équation 2.14). Cette mesure ne considère que les documents qui sont explicitement évalués, alors que d'autres mesures supposent généralement que les documents non jugés sont sans pertinence. bpref est calculé comme suit :

$$\text{bpref} = \frac{1}{|R|} \sum_{r \in R} \left(1 - \frac{|\forall n(n \in \bar{R} \wedge n < r)|}{|R|} \right) \tag{2.14}$$

où r représente un document dans l'ensemble des documents pertinents noté R, n est un document non pertinent dans l'ensemble de documents non pertinents $\bar{R}$, tel que n se produit avant r dans la liste classée.
Les documents qui n'ont pas été évalués en fonction de leur pertinence n'affectent donc pas la mesure.

2.7.2 Campagnes de test et évaluation

Conférence Text REtrieval (TREC)

La méthodologie d'évaluation et les mesures décrites dans la partie précédente sont au cœur de la campagne d'évaluation de la Conférence Text REtrieval (TREC) [Voorhees et Harman, 2005]. TREC vise à fournir une plateforme commune pour évaluer les SRI en développant des collections d'évaluation. Une collection de test est composée d'un corpus de documents, d'un ensemble de requêtes (souvent appelées rubriques) et d'un ensemble de jugements de pertinence fournis par des utilisateurs experts. Le corpus documentaire et les requêtes associées sont mis à la disposition des équipes participant à la campagne. Les équipes utilisent la méthode qu'elles ont élaborée pour exécuter les requêtes et soumettent leurs résultats, sous la forme d'une liste classée de documents, aux organisateurs du TREC. Les organisateurs évaluent ensuite la contribution de chaque équipe en fonction des jugements de pertinence. Les premières collections de test comportaient un petit nombre de documents. Par exemple, les collections d'articles ACM (CACM) ne contiennent que 3 024 documents. De petites collections de documents peuvent être évaluées complètement par des juges experts. Toutefois, à mesure que les collections de documents se sont développées - la collection ClueWeb contient 1,2 milliard de documents Web - il est devenu impossible d'évaluer tous les documents. Pour faire face à cette question, TREC a utilisé des techniques de mise en commun pour sélectionner un sous-ensemble approprié de documents pour évaluation par des experts. Le regroupement se fait en prenant un échantillon de documents pour chaque requête de chaque équipe participante. Ces documents sont fusionnés en un ensemble unique (appelé le pool), qui est ensuite fourni aux évaluateurs experts. Si des systèmes suffisamment divers contribuent au pool, un sous-ensemble représentatif du document sera évalué et les jugements de pertinence ne devraient pas favoriser un système particulier. TREC est organisé en sous-défis distincts, appelés *Tracks*, qui se concentrent

sur des applications de RI particulières (par exemple le Web Track est spécifique à la recherche de documents Web). En 2011, TREC a lancé le MedTrack (Medical Records Track), conçu pour « encourager la recherche sur la technologie permettant de récupérer les dossiers de santé électroniques sur la base du contenu sémantique des champs de texte libre » [Voorhees et Hersh, 2012]. La collection de documents utilisée dans TREC MedTrack comprenait 100 866 documents issus de dossiers cliniques désidentifiés provenant d'hôpitaux américains. Les thèmes et jugements de pertinence ont été créés par les médecins, les sujets reflétant les types de requêtes qui pourraient être utilisés pour identifier les patients admissibles à l'inclusion dans les essais cliniques.

Conference and Labs of the Evaluation Forum (CLEF)

La conférence CLEF se compose d'une conférence indépendante évaluée par des pairs sur un large éventail de questions dans les domaines de l'évaluation de l'accès à l'information multilingue et multimodale et d'un ensemble de laboratoires et d'ateliers conçus pour tester différents aspects des systèmes de recherche d'information mono et inter langues. Elle propose plusieurs campagnes d'évaluation de la recherche d'information, en particulier basées sur l'expérimentation et les tâches partagées : CLEF Labs, TREC, NTCIR, FIRE, MediaEval, RomIP, SemEval, TAC. L'une de ces campagnes, CLEF eHealth, est dédiée au développement de méthodes de traitement et de ressources dans un contexte multilingue pour enrichir le texte libre numérisé dans le domaine de la santé. La campagne propose plusieurs tâches liées aux problématiques de recherche d'information et d'extraction d'information dans le domaine de la santé. En 2016, trois tâches ont été étudiées : l'extraction de l'information de transfert liée aux changements de soins infirmiers australiens, l'extraction de l'information dans des corpus cliniques francophones et enfin la recherche d'information multilingue axée sur le patient en tenant compte des variations de requête [Kelly et al., 2016].

Campagnes de test i2b2

La vaste adoption de la plateforme i2b2 permet l'organisation annuelle de campagnes de test dédiées particulièrement aux tâches d'extraction de l'information dans les documents cliniques. Des corpus spécifiques sont diffusés et leur annotation est évaluée de façon indépendante. Par exemple, la campagne 2012 se concentrait sur les relations temporelles dans les comptes-rendus cliniques. Un corpus de résumés de décharge annotés avec des informations temporelles, utilisés pour le développement et l'évaluation de systèmes de raisonnement temporel était fourni. L'évaluation portait sur l'annotation (1) d'événements cliniquement significatifs, y compris les concepts cliniques tels que les problèmes, les tests, les traitements et les services cliniques, ainsi que les événements pertinents pour l'historique clinique du patient, tels que les admissions, les transferts entre les départements, etc. (2) les expressions temporelles, en

se référant aux phrases de dates, heures, durées ou fréquences dans le texte clinique et (3) les relations temporelles, entre les événements cliniques et les expressions temporelles [Sun et al., 2013]. La campagne 2016 se concentrait sur les problématiques de dé-identification des textes et comptes-rendus cliniques ainsi que l'annotation de termes psychiatriques.

2.8 Synthèse

Dans cette partie, nous avons défini la notion d'information en santé et décrit les ressources disponibles ainsi que leurs représentations dans différents SOC. Plusieurs de ces SOC seront exploités plus loin dans cette thèse, à la fois dans la modélisation de données pour l'intégration de données clinomiques dans les DPI et dans l'indexation de textes médicaux à l'aide de vocabulaires contrôlés : l'UMLS, la CIM10, la SNOMED CT et le MeSH pour citer les principaux ou encore la Gene Ontology plus spécifique à la problématique de gestion des données clinomiques.

L'organisation des données au sein des DPI et d'entrepôts de données dédiés à la recherche clinique et translationnelle a également été abordée. Les systèmes disponibles et décrits ici ont pour point commun de proposer uniquement des fonctionnalités sur des cohortes de patients tandis que la solution développée dans le cadre de cette thèse permet la manipulation de données cliniques et omiques à l'échelle d'un ou plusieurs patients.

Enfin, nous avons décrit les différents méthodes et modèles d'indexation et de recherche d'information applicables au domaine de la santé. Dans cette thèse, nous nous concentrerons sur l'indexation à l'aide de vocabulaires contrôlés appliquée à des textes médicaux. La structuration de ces informations est en effet un verrou important pour une RI efficace, dans les DPI mais aussi d'autres textes comme des certificats de décès exploités dans cette thèse. L'organisation des connaissances est ainsi centrale dans cette thèse.

Chapitre 3

Modélisation de données omiques et cliniques pour le Dossier Patient Informatisé

Sommaire

Dans ce chapitre, il s'agit de présenter les différentes étapes ayant permis la conception d'un modèle générique des données cliniques et omiques. Le développement de la partie clinique du modèle réalisé par l'équipe du D2IM du CHU de Rouen s'est déroulé dans le cadre du projet RAVEL. Le développement de la partie omique dont j'avais la tâche a nécessité plusieurs étapes décrites ici : d'abord le recensement et l'identification des différents types de données omiques, ensuite la collecte de données auprès de divers acteurs et enfin la modélisation de ces données. La conception de ce modèle implique ainsi la coordonnation de types de données hétérogènes : les données cliniques et administratives du patient, essentiellement textuelles et numériques, et des données de biologie moléculaire.

3.1 L'intégration de données hétérogènes

L'intégration de données hétérogènes est une problématique grandissante avec l'expansion rapide des connaissances biomédicales, la réduction des coûts informatiques et la diffusion de l'accès à Internet. Aujourd'hui, les bases de données à travers le monde contiennent des données biomédicales allant des résultats cliniques d'un patient individuel à la structure génétique de notre espèce. Beaucoup de ces systèmes sont accessibles via Internet. Dans l'ensemble, ces données englobent de l'information et des connaissances qui peuvent améliorer considérablement la recherche fondamentale, les soins aux patients et la santé publique. Cependant, le volume et la disponibilité exceptionnels de ces données ont augmenté grâce à un processus largement décentralisé qui a permis aux organisations de répondre à des besoins de données spécifiques ou locaux sans les obliger à coordonner et à normaliser leurs implémentations de base de données. Ce processus a abouti à une mosaïque hétérogène de sources de données, rendant leur accès et leur agrégation difficiles d'un point de vue pratique [Hamid et al., 2009].

Ainsi, de plus en plus, l'analyse de données exige l'utilisation de données générées et gérées par des tiers. La variété des données utilisées, autant que l'échelle des données analysées, peut ainsi s'avérer un facteur limitant dans les efforts d'analyse des données. Avec l'avènement du traitement de données de masse en sciences, les données peuvent provenir à des échelles significatives à partir d'une seule machine (comme les séquenceurs haut débit) ou peuvent être réparties entre scientifiques dans un projet distribué à grande échelle. Dans ces cas, de plus en plus les communautés développent des vocabulaires contrôlés ou des ontologies et des normes de métadonnées pour les efforts d'intégration de données. La possibilité de fédérer des données dans des ensembles de données, des catalogues et des domaines peut fournir aux utilisateurs la possibilité de trouver, d'accéder, d'intégrer et d'analyser des combinaisons d'ensembles de données en fonction de leurs besoins.

Plusieurs grandes étapes nécessaires à l'intégration de données hétérogènes peuvent

être identifiées : la découverte et la collecte des données, l'intégration des données à proprement parler et leur validation [Hendler, 2014]. La découverte de données pertinentes à un projet est la première étape réelle. La recherche de données est complexe, même dans un entrepôt, mais à l'extérieur d'un entrepôt, et sur le Web, elle est encore plus difficile. Plusieurs approches explorent la façon dont divers types de métadonnées peuvent être utilisées pour faciliter l'exploration des données comme la recherche facettée ou l'utilisation de métadonnées plus précises voire spécifiques au domaine. Par exemple, pour les données de recherche scientifique, la Research Data Alliance [1] a créé un certain nombre de groupes de travail qui explorent le développement de normes de métadonnées pour des domaines scientifiques spécifiques. D'autres approches ont exploré l'extraction de métadonnées à partir de descriptions de données ou de données structurées sur des pages Web [Cafarella et al., 2011].

L'intégration de données peut s'envisager sous plusieurs formes. Le schéma relationnel traditionnel centré autour du concept d'une table (ou relation) qui se compose de lignes et de colonnes (attributs). Le modèle relationnel est une méthode bien connue et robuste de représentation des données, mais l'une de ses principales critiques est que la modélisation d'objets structurés, de hiérarchies, tel que les entités biologiques, n'est pas immédiatement intuitive pour quiconque, à l'exception des concepteurs [Stein, 2003]. L'utilisation de données semi-structurées libère de la structure rigide du modèle de données relationnelles [Achard et al., 2001a]. Les données semi-structurées sont essentiellement des données avec une série d'étiquettes et de valeurs associées. XML est un format recommandé par le WWW Consortium pour l'échange de données sur le Web et est parfaitement adapté pour décrire les données semi-structurées [Abiteboul et al., 2000]. En termes de représentation des données et connaissances, l'utilisation d'ontologies présente l'avantage supplémentaire de faciliter l'interaction entre les chercheurs dans différents domaines du savoir et de permettre l'interopérabilité entre les bases de données et les programmes, qui sont essentiels au travail collaboratif en sciences [Bodenreider, 2008].

Enfin, La validation des données vise à mettre en évidence certains des problèmes largement rencontrés lors de l'intégration de données hétérogènes : les données manquantes, les données saisies incorrectement ou d'autres problèmes communs de "données sales". Souvent, l'intégration des données permet de mettre en évidence ces problèmes qui n'étaient pas visibles lorsque les données étaient étudiées individuellement [Hendler, 2014].

1. https://rd-alliance.org/

3.2 Le projet RAVEL

Aujourd'hui, avec le développement des DPI, la quantité de données cliniques augmente constamment et ces données deviennent finalement accessibles aux patients. Les DPI contiennent une grande variété de données cliniques hétérogènes qui présentent différentes valeurs d'information et décisionnelles pour les professionnels de la santé : elles sont utilisées de différentes façons et dans différentes situations. En outre, une grande quantité de données sont encore enregistrées sous forme de documents non structurés et narratifs, bien que ces données contiennent des informations cruciales dans un processus de soins de santé plus efficace, telles que des notes cliniques, des observations, des rapports récapitulatifs pour n'en nommer que quelques-uns. La simple accumulation de données cliniques dans les DPI, ainsi que leur diversité, entraîne le risque de débordement d'information tant pour les praticiens de la santé que pour les patients. En outre, on estime que jusqu'à 50% de l'information clinique nécessaire pour décrire le séjour hospitalier des patients n'est disponible que dans des documents descriptifs non structurés dans les dossiers des patients [Turchin et al., 2006]. Ainsi, l'évolution des dossiers des patients vers le DPI entraîne des paradoxes : les données sont exhaustives, mais peu exploitées et ne proposent pas de présentation structurée et hiérarchisée, ce qui freine le processus décisionnel par les médecins. Il est donc essentiel de fournir des outils de recherche et de visualisation robustes et efficaces pour les données DPI. Dans ce cadre, le projet RAVEL étudie et met en œuvre : (i) les méthodes d'indexation les plus récentes afin d'enrichir les données des DPI sémantiquement structurées et non structurées. Cet enrichissement sera utile pour (ii) l'optimisation de la récupération de l'information, et (iii) la visualisation des données des patients.

3.2.1 Le modèle de données RAVEL

L'utilisation d'un DPI est devenue une pratique courante dans tous les hôpitaux. Les modèles de ces DPI varient d'un hôpital à l'autre, mais ils sont le plus souvent complexes. À titre d'exemple, le DPI du CHU de Rouen contient plus de 100 entités. Le modèle de données RAVEL décrit comment sont représentées les données cliniques dans la base de données. Il s'appuie sur une base de données relationnelle afin de stocker les données cliniques. Il est basé sur un modèle conceptuel générique [Achard et al., 2001b] compact, optimisé pour la recherche d'information. Il est composé de onze entités dans lesquelles sont réparties les informations concernant les séjours, analyses et actes médicaux, prescriptions et informations civiles. La Figure 3.1 représente ces entités et leurs attributs et relations.

Ce modèle de données RAVEL est intégré au méta modèle utilisé dans le système d'information CISMeF (voir la section 4.1.2). L'équipe CISMeF a conçu un modèle logique générique de données pouvant encapsuler n'importe quel modèle. Ce modèle se

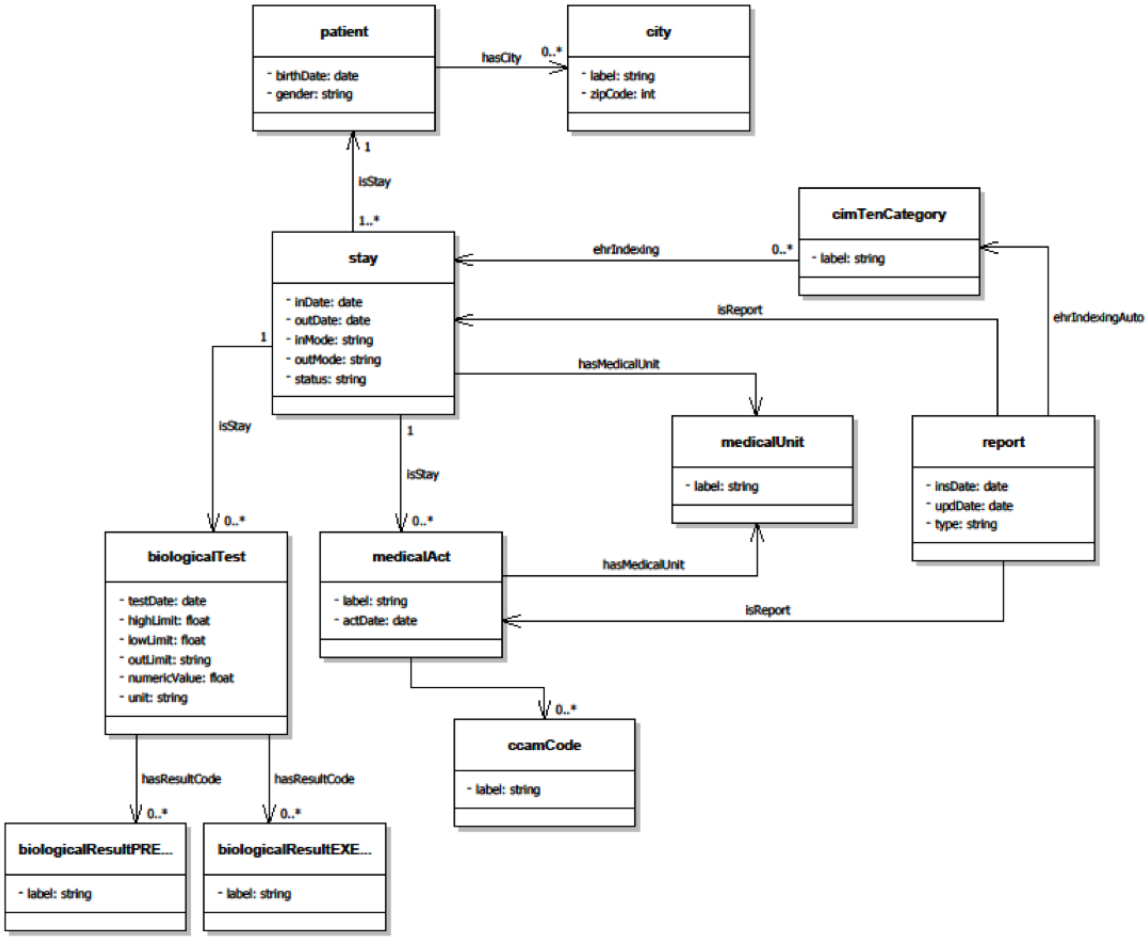

FIGURE 3.1 – Modèle physique des données RAVEL.

compose en deux parties : d'une part le « modèle » pour définir un modèle conceptuel de données et d'autre part l'instance du modèle pour stocker les données elles-mêmes. Dans le système d'information CISMeF, ce méta modèle intègre l'ensemble des ressources terminologiques et des documents indexés.

3.2.2 Données cliniques

Actuellement, le CHU de Rouen utilise le logiciel C Page Dossier Patient (CDP) pour la gestion des dossiers patients informatisés (environ 1 800 000).

Parmi ces dossiers, 2 021 ont d'abord été sélectionnés selon des critères médicaux (spécialités et complexité) puis dé-indentifiés pour être intégrés à RAVEL. Ces dossiers comportent les données concernant (i) le patient (âge, sexe), (ii) les séjours, (iii) les actes médicaux, (iv) les analyses biologiques/imagerie, (v) les prescriptions. La volumétrie est de 182 000 séjours, 284 000 actes, 286 0000 analyses biologiques, 152 000 comptes-rendus pour 2 021 patients.

Chaque dossier est codé par des terminologies médicales comme la CIM10 ou la CCAM grâce au codage Programme de Médicalisation des Systèmes d'Information (PMSI), réalisé manuellement par des experts (le plus souvent des médecins) après chaque hospitalisation. Le but de ce programme est à l'origine de la facturation des actes et du suivi des patients (visée socio-économique).

3.3 Collecte de données omiques

3.3.1 Bases de connaissances

Afin de disposer d'informations de référence pour décrire des données omiques expérimentales, plusieurs bases de données internationales ont été utilisées et certaines de leurs informations intégrées au modèle de données omiques.

Gènes : NCBI Gene, Gene Ontology

Pour la description des gènes, la banque Gene du National Center for Biotechnology Information (NCBI) a été utilisée [Maglott et al., 2007]. Cette base de données recueille des informations validées sur les gènes non prédits de diverses espèces, dont l'homme. Ces informations sont exhaustives, on trouve notamment le nom du gène, ses coordonnées sur le génome, son type (codant, pseudogènes...) ou encore un résumé descriptif. L'intégralité des données est accessible sous forme d'un fichier binaire, convertible en XML, et est mise à jour quotidiennement.

Ces informations ont été complétées des annotations issues de la GO associées aux gènes, déjà intégrées dans le système d'information de CISMeF. La GO est un projet bioinformatique destiné à structurer la description des gènes et des produits géniques dans le cadre d'une ontologie commune à toutes les espèces [The Gene Ontology Consortium, 2008].

Protéines : Uniprot K/B

Ensuite, concernant les protéines, la banque UniprotKB [UniProt Consortium, 2013] a été utilisée afin de servir de référence en particulier aux analyses d'expression protéique. Cette base de données comporte une collection d'informations fonctionnelles sur les protéines telles que le nom et la description de la protéine, sa fonction ou encore sa localisation cellulaire. UniprotKB comporte deux sections : l'une, UniprotKB/Swiss-Prot, est manuellement annotée et supervisée, l'autre UniprotKB/TrEMBL n'est pas supervisée et uniquement annotée. Seule UniprotKB/Swiss-Prot a été utilisée dans notre modèle de données. L'intégralité des données d'UniprotKB est accessible par un fichier XML et est mise à jour régulièrement (environ tous les mois).

barcode	chromosome	start	stop	num.mark	seg.mean
TCGA-02-0001-01C-01D-0185-02	1	554267	72533855	6384	0.0456
TCGA-02-0001-01C-01D-0185-02	1	72550247	72568008	2	1.2471
TCGA-02-0001-01C-01D-0185-02	1	72602596	74674719	93	0.0994
TCGA-02-0001-01C-01D-0185-02	1	74693651	74877529	20	-0.3621
TCGA-02-0001-01C-01D-0185-02	1	74885003	74952060	7	-0.6845
TCGA-02-0001-01C-01D-0185-02	1	74961517	75110250	10	-0.3235
TCGA-02-0001-01C-01D-0185-02	1	75148401	116948645	3057	0.0836
TCGA-02-0001-01C-01D-0185-02	1	116950995	117008722	5	-0.4321
TCGA-02-0001-01C-01D-0185-02	1	117014630	150816179	803	0.0395
TCGA-02-0001-01C-01D-0185-02	1	150823072	150848508	4	-1.5222

FIGURE 3.2 – Exemple de fichier de données issu de la base de données TGCA.

Phénotypes : OMIM

Enfin, les ressources du catalogue Online Mendelian Inheritance in Man (OMIM) ont également été intégrées dans le but de disposer d'informations de référence concernant les maladies génétiques [HAMOSH et al., 2005]. Les données de description de gènes incluses dans le catalogue OMIM n'ont pas été utilisées puisque redondantes avec la base NCBI Gene, seules les données correspondant à la description des phénotypes ont été considérées. Elles complètent les données d'Orphanet, une base de données concernant les maladies orphelines, également intégrée au système d'information CISMeF.

3.3.2 Données expérimentales

Pour appuyer la modélisation des données omiques, j'ai préalablement collecté divers types de données expérimentales, afin de les étudier et par la suite de peupler la base de données.

The Genome Cancer Atlas

Les données issues du portail The Genome Cancer Atlas (TGCA) ont été principalement utilisées. Le portail TGCA est un projet débuté en 2005 dont l'objectif est de cataloguer les mutations génétiques responsables de cancers, en utilisant les techniques d'analyse omiques et la bioinformatique. Le projet, supervisé par le National Institutes of Health (NIH) américain, a la particularité de proposer des cohortes de patients importantes ainsi qu'une grande diversité d'études omiques et de techniques différentes couvrant 33 types de cancers et représentant plus de 2 pétabytes de données. Les données sont disponibles au format texte tabulé (voir FIGURE 3.2).

Variations génétiques

Dans le cadre de la collaboration avec plusieurs laboratoires de recherche, plusieurs jeux de données omiques ont été obtenus. Le laboratoire INSERM U918 Génétique

TABLEAU 3.1 – Les différents types de données en sciences omiques.

Type de données	Technologie utilisée
Analyse du nombre de copies	Altération du nombre de copies pour une région segmentée par échantillon
Méthylation de l'ADN	Valeurs beta calculées pour une région génomique par échantillon
Expression : exon	Signal d'expression normalisé par exon par échantillon
Expression : gène	Signal d'expression normalisé par gène par échantillon
Expression : miRNA	Signal d'expression normalisé par miRNA par échantillon
Expression : jonction	Signal d'expression normalisé par jonction par échantillon
Expression : transcrit	Signal d'expression normalisé par transcrit par échantillon
Expression : protéine	Signal d'expression normalisé par protéine par échantillon
Gènes de fusion	Signal d'expression normalisé par protéine par échantillon
Variants	Variants validés par échantillon

et Clinique des proliférations Lyphoïdes du Centre H. Becquerel, dirigé par le Pr. F. JARDIN, nous a fourni les données d'analyse Comparative Genomic Hybridization (CGH) de 20 patients. L'équipe du Dr. D. CAMPION (INSERM 614 Génétique du cancer et des maladies neuropsychiatriques, dirigé par le Pr. T. FRÉBOURG) nous a permis d'utiliser les données de séquençage d'exomes de 120 patients atteints de la maladie d'Alzheimer.

3.4 Recensement des types de données omiques

Préalablement à la conception du modèle de données omiques, il a été nécessaire de réaliser une revue exhaustive des différents types de données que devait prendre en charge le modèle de données. Une vue d'ensemble de ces différents types de données s'avère en effet nécessaire afin de développer un modèle générique et flexible, seule solution nous semblant adaptée à l'hétérogénéité des données omiques.

3.4.1 Les différents types de données omiques

Les sciences omiques ciblent différents objectifs et génèrent ainsi différents types de données. Il peut s'agir de détection de variants ou d'analyse d'expression de segments génomiques tels que des gènes, exons ou encore de protéines. Afin de cerner les différentes données générées par les scientifiques ainsi que leurs besoins, j'ai été amenée à rencontrer plusieurs acteurs de la recherche utilisant couramment les techniques d'analyse omique en laboratoire.

Données quantitatives

Dans le cas de données de quantification, il s'agit dans la plupart des cas d'une valeur numérique, représentant le taux d'expression d'une entité (sonde, gène, protéine, exon...) ou encore le taux de méthylation de l'ADN. Ces valeurs peuvent être déterminées par des technologies de puces à ADN ou de séquençage.

Chacune de ces entités peut présenter des caractéristiques différentes. Globalement, on peut discriminer (i) les gènes, (ii) les protéines et (iii) les segments caractérisés principalement par leur localisation sur le génome (variants structuraux, exons, transcrits, variants d'épissage). Ces types sont décrits dans le Tableau 3.1.

Variants génétiques

Les variants tels que les Single Nucleotid Polymorphism (SNP) et petites insertions délétions (indels) mais également les variants structuraux ou Structural Variant (SV) peuvent être représentés par des chaînes de caractères formalisées. La nomenclature Human Genome Variation Society (HGVS) permet en effet de nommer les différents types de variations en fonction de leur localisation. Ces variants peuvent être identifiés par des techniques de puces à ADN ou également de séquençage.

Les variants structuraux peuvent être de plus caractérisés par la valeur d'un signal correspondant à l'analyse de sondes réparties sur le génome par la technique de CGH. Cette valeur représentera par exemple la délétion ou l'insertion d'un segment. Les variants du nombre de copies ou Copy Number Variation (CNV) seront également caractérisés par une valeur numérique représentant le nombre de copies.

3.4.2 Niveaux d'interprétation des données omiques

L'une des problématiques importantes pour la conception du modèle de données est de sélectionner les informations pertinentes dans le cadre défini par l'intégration avec des données cliniques. En effet, les sciences omiques génèrent des volumes de données très importants et ces données doivent être traitées et interprétées pour être exploitables à l'échelle d'un patient ou d'un groupe de patients. J'ai ainsi réalisé une catégorisation des données omiques en quatre niveaux d'interprétation, des données brutes aux données recoupées et interprétées en régions d'intérêt. Ces niveaux sont décrits sur le Tableau 3.2.

Les données brutes (niveau 1) correspondent aux données non normalisées. Pour un séquençage, ce niveau correspond aux données brutes sorties du séquenceur. Elles peuvent être accessibles par des fichiers textes ou binaires, dont le format dépendra fréquemment du matériel utilisé. Le plus souvent, le volume de données est très important (jusqu'à plusieurs giga-octets pour une analyse) et ces données ne peuvent être interprétées manuellement.

	Type	Description	Exemple
1	Données brutes	Données non normalisées	— Fichier BAM (données de séquences) — Fichier CEL Affymetrix (données de microarrays)
2	Données traitées	Données normalisées pour un échantillon	— Signal d'une sonde ou d'un groupe de sondes par échantillon — Amplification/délétion d'une sonde dans un échantillon — Mutation supposée pour un échantillon
3	Données interprétées	Données agrégées pour un échantillon	— Signal d'expression d'un gène pour un échantillon — Amplification/délétion d'une région segmentée dans un échantillon — Mutation validée pour un échantillon
4	Régions d'intérêt	Associations quantifiées et croisées entre différents types d'échantillons. Associations basées sur : — Anomalies moléculaires — Caractéristiques de l'échantillon — Variables cliniques	— Une région génomique est amplifiée dans 10% des gliomes.

Tableau 3.2 – Les différents types de niveaux d'interprétation des données omiques.

Les données traitées (niveau 2) correspondent aux données normalisées par une méthode statistique comme la LOWESS par exemple. Il s'agit du signal d'une sonde ou d'un groupe de sondes pour une analyse d'expression, ou encore d'un variant supposé pour un échantillon. Ces données sont accessibles par des fichiers textes sur des banques de dépôt comme Gene Expression Omnibus (GEO) ou ArrayExpress. Le volume de données est réduit, mais reste important. Pour une analyse d'expression de gènes, le fichier de résultats concernant un seul échantillon peut aller jusqu'à une centaine de mega-octets. L'interprétation manuelle reste délicate, l'information concernant des sondes ou des variants non validés.

Les données interprétées (niveau 3) regroupent des données qui ont été agrégées pour un échantillon. Par exemple, pour l'analyse d'expression de gènes, il s'agira du signal d'expression d'un gène, les signaux des sondes correspondant à ce gène ayant été agrégés ou encore d'un variant validé. Ce type de données est disponible en fichier texte, le plus souvent tabulé.

Cependant, il n'existe pas de standard établi. Le volume de données est dans ce cas réduit, un fichier de résultats pour une analyse d'expression peut représenter de quelques kilo-octets jusqu'à 1 Mo, selon le nombre de gènes analysés.

Ce niveau de données n'est pas accessible dans les banques de dépôt ArrayExpress et GEO, qui ne proposent que des données de niveau 1 et 2. Peu de banques de données proposent ces données interprétées. Le portail TGCA offre les données recueillies dans une vingtaine d'études impliquant jusqu'à plusieurs centaines de patients. Les techniques utilisées sont variées et couvrent tous les types de données vus précédemment.

Dans ce cas, on dispose d'informations validées et exploitables. Ce niveau de données paraît donc pertinent à intégrer dans un dossier médical. Le Tableau 3.3 présente pour chaque type de données l'information de niveau 3 correspondante.

Enfin, les données interprétées et agrégées correspondent au niveau d'interprétation le plus élevé (niveau 4). Il s'agit de réaliser des associations quantifiées et croisées entre différents types d'échantillons afin d'isoler une région d'intérêt. Cette interprétation approfondie des données omiques nécessite une expertise biostatistique et biologique humaine pointue. L'aboutissement à ce niveau d'interprétation est notamment l'un des buts de la plateforme tranSMART. Très peu de ressources sont disponibles de façon standardisée et formalisée. De plus, de telles données ne s'appliquent plus à l'échelle du patient, mais à celle du phénotype, il n'est donc pas adapté à l'intégration avec des données cliniques. Cependant, les régions d'intérêt isolées représentent une information pertinente, notamment à des fins de diagnostic ou de recherche.

Ainsi, le développement du modèle de données omiques est basé sur les données omiques de niveau 3. En effet, elles semblent les plus adaptées à figurer dans un dossier médical. Elles présentent directement une information exploitable telle qu'un variant validé ou la délétion d'un segment génomique. Elles sont donc les plus pertinentes, à

la fois pour la visualisation dans le dossier médical et pour la recherche d'information.

Type de données	Niveau 3 : Description
Variants structuraux	Altération d'une région segmentée par échantillon
Analyse du nombre de copies	Altération du nombre de copies pour une région segmentée par échantillon
Méthylation de l'ADN	Valeurs bêta calculées pour une région génomique par échantillon
Expression : exon	Signal d'expression normalisé par exon par échantillon
Expression : gène	Signal d'expression normalisé par gène par échantillon
Expression : miRNA	Signal d'expression normalisé par miRNA par échantillon
Expression : jonction	Signal d'expression normalisé par jonction par échantillon
Expression : transcrit	Signal d'expression normalisé par transcrit par échantillon
Expression : protéine	Signal d'expression normalisé par protéine par échantillon
Variants (SNP, indels)	Variants validés par échantillon

TABLEAU 3.3 – Description du niveau d'interprétation 3 pour les principaux types de données omiques sélectionnés pour concevoir le modèle de données.

3.5 Modèle de données omiques

Le modèle de données omiques se compose de trois parties principales (voir FIGURE 3.3) gérant (i) les informations concernant l'étude réalisée et le laboratoire responsable, (ii) les données de variants de type SNP/indels et (iii) les données d'expression ou globalement de quantification [CABOT et al., 2014].

3.5.1 Gestion des études omiques

Cette partie gère les informations concernant l'étude omique réalisée. L'entité `OMI_LAB` regroupe les informations concernant le (ou les) laboratoire(s) ayant conduit l'étude. Ensuite, les informations concernant les responsables de l'étude sont représentées par l'entité `OMI_SUBMITTER`. Les données liées à l'étude telles que son nom, l'équipement

utilisé ou le protocole sont représentées par l'entité `OMI_STUDY`. Dans certains cas, en particulier si l'étude porte sur une analyse d'expression, les valeurs numériques obtenues peuvent être décrites par une unité de mesure. Cette information correspond alors à l'entité `OMI_UNITS`, accompagnée de sa signification.

3.5.2 Gestion des données d'expression et de quantification

L'objet de l'analyse quantitative peut être de trois types : (i) un gène, (ii), une protéine, (iii) tout segment caractérisé par une position de début et de fin. Ces informations sont représentées respectivement par les entités `GENE`, `PROTEIN` et `OMI_SEGMENT`.

L'entité `GENE` représente les données de l'intégralité des gènes et des microARN identifiés chez l'Homme, issues de la base de données Gene du NCBI. Chaque gène est associé à ses annotations GO, ainsi qu'aux maladies références dans OMIM qui lui sont liées. De la même manière, l'entité `PROTEIN` contient les informations concernant l'intégralité des protéines isolées chez l'Homme, issues de la base de données UniprotKB/Swissprot. Chaque entrée est également liée aux annotations GO et aux maladies OMIM ou Orphanet associées. Pour chaque analyse d'expression de gènes ou de protéines, les données utilisées sont donc issues de ces deux bases de données exhaustives et supervisées, régulièrement mises à jour, afin de garantir d'une part la qualité des données et d'autre part d'accélérer l'intégration de données expérimentales.

En effet, la disponibilité des données de gènes et de protéines dans notre base de données permet d'éviter l'intégration progressive de ces informations, au fur et à mesure de l'intégration de données expérimentales.

L'entité `SEGMENT` est générique et peut représenter les données d'exons, de transcrits ou encore de variants de structure (larges insertions et délétions). Chaque segment est lié au(x) entité(s) `GENE` correspondante(s), selon les positions de début et de fin, et du chromosome. Chaque segment est ainsi également associé aux protéines correspondantes, aux termes GO du (des) gène(s) ainsi qu'aux données de pathologies OMIM ou Orphanet associées au(x) gène(s).

La valeur numérique associée à chaque segment, gène, ou protéine analysés est stockée dans la table `OMI_QUANTIF`. Chaque entrée est accompagnée de clés étrangères liant l'étude et le patient (modèle RAVEL) concernés. Les données stockées sont la valeur brute obtenue lors de l'analyse ainsi que la valeur normalisée, l'interprétation de cette valeur (une délétion ou insertion par exemple dans le cas d'une analyse de variants structuraux) et enfin éventuellement un commentaire.

3.5.3 Gestion des données de variants

Les informations concernant les variants de type SNP et indels sont représentées par l'entité `OMI_VAR`. Ces données comportent notamment les informations telles que le

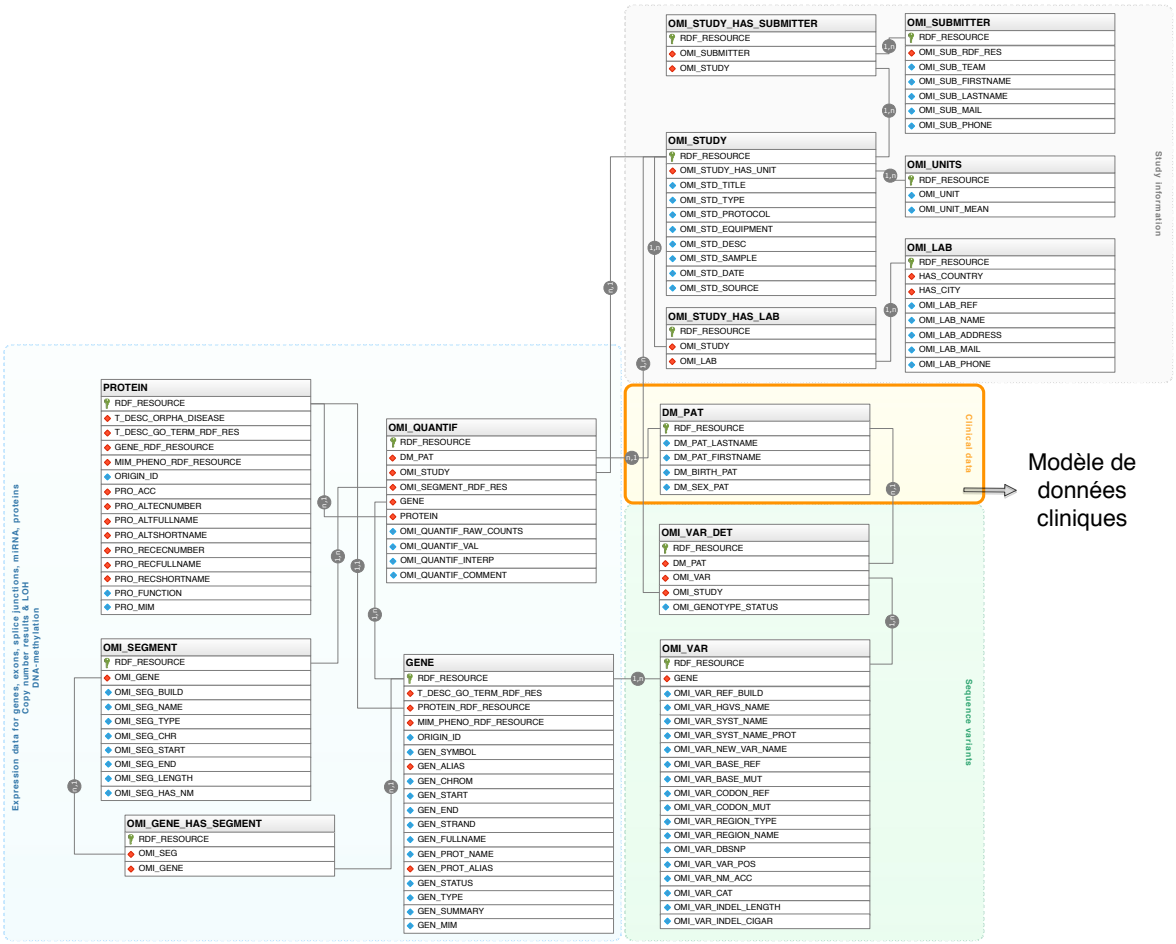

FIGURE 3.3 – Modèle logique des données omiques.

nom de la variation selon les différentes conventions, sa localisation, la base de référence et la base mutée dans le cas d'un SNP. Ces informations sont reliées à plusieurs autres entités. D'une part une association est faite avec le patient grâce à la table d'association `OMI_PAT_HAS_VAR`, cette association fait alors le lien avec le modèle RAVEL des données cliniques. Pour chaque entité, une association est également existante avec l'étude. Ainsi, une même variation apparaissant chez plusieurs patients ne sera stockée qu'une seule fois, évitant toute redondance des données de variations.

Enfin, chaque variation est également reliée au gène correspondant. Cette association permet de relier chaque variation non seulement au gène, mais également aux maladies OMIM ou Orphanet référencées pour ce gène, ou encore ses annotations GO.

3.6 Synthèse

La conception et le développement d'un système de recherche d'information et de visualisation des données omiques et cliniques au sein d'un DPI nécessite en premier lieu un modèle de données générique gérant à la fois des données cliniques et omiques.

Le premier objectif de ce travail débuté dans le cadre de mon mémoire de master a ainsi été le recensement et l'identification des différents types de données omiques à modéliser. Ce recensement des données et la compréhension des différentes étapes de leur analyse a mené à une classification des types de données omiques. Ce recensement a également permis d'identifier des acteurs clés pour d'une part regrouper des informations et connaissances servant l'exploitation ultérieure des données et d'autre part mettre au point des jeux de données expérimentales en vue de leur utilisation dans l'application prototype. Ce travail m'a permis d'aboutir à un modèle de données omiques générique, compact et évolutif capable de gérer la vaste majorité des données expérimentales aujourd'hui utilisés en recherche clinique et les informations de référence produites par la communauté scientifique qui leur sont liées.

Chapitre 4

Recherche d'information dans les données cliniques et omiques au sein du Dossier Patient Informatisé

Sommaire

Ce chapitre présente les développements réalisés pour la conception d'une application prototype permettant la visualisation de données cliniques et omiques et la mise au point d'un outil de recherche dans ces données. On peut distinguer trois grandes parties. Tout d'abord, le premier objectif est l'intégration des données de référence et des données expérimentales sur la base du modèle de données conçu décrit dans le chapitre précédent. La seconde étape est la conception de vues permettant la visualisation des données omiques dans l'application prototype et l'adaptation du moteur de recherche existant pour les données cliniques pour permettre l'interrogation conjointe des données cliniques et omiques, à l'échelle d'un ou plusieurs patients. Enfin, l'évaluation de notre solution est réalisée selon deux axes : l'analyse comparative avec les plateformes de référence que sont i2b2 et TranSMART et l'évaluation de la partie clinique de notre système dans le cadre du projet RAVEL à travers un cas d'usage.

4.1 Intégration de données cliniques et omiques

4.1.1 Choix des ressources à intégrer

Le modèle de données développé est ainsi en partie basé sur l'intégration de données de références externes, concernant les gènes et protéines ainsi que les pathologies associées.

L'intégration de ces données externes a plusieurs objectifs. D'une part, la disponibilité en local de l'ensemble des gènes et protéines permet d'éviter l'interrogation de serveurs distants lors du traitement de données omiques expérimentales. D'autre part, l'accessibilité de ces données peut se révéler utile pour la recherche d'information.

Enfin, ces données sont ainsi également accessibles sur le portail HeTOP développé par l'équipe, qui met à disposition un grand nombre de terminologies de santé dans plusieurs langues, et ce, dans plusieurs buts : aider à l'indexation (codage, annotation), accéder à des ressources bibliographiques, enseigner/apprendre la médecine ou encore réaliser des audits de terminologies. Ce portail est décrit dans la section 5.1.1.

4.1.2 Intégration de terminologies et ontologies au sein du méta-modèle 3M

Le méta-modèle 3M

L'interopérabilité sémantique au sein du système d'information CISMeF repose sur un modèle unifié de vocabulaires nommé méta-modèle 3M développé dans le cadre de la thèse de Julien Grosjean [Grosjean, 2014]. Le méta-modèle 3M est un modèle commun à tous les termes, quelle que soit la terminologie à laquelle ils appartiennent. Il peut être considéré comme un méta-modèle ou une ontologie supérieure conçue pour

supporter une interopérabilité sémantique générale entre les terminologies qui le composent. Le méta-modèle 3M fournit des propriétés d'alignement telles que celles définies dans le langage de la ressource considérée. Le méta-modèle est fondamentalement multilingue car les termes préférés, les synonymes ou d'autres attributs textuels peuvent être définis par un code de langue (« en » pour l'anglais, « fr » pour le français, etc.). Chaque terminologie incluse dans HeTOP est un enrichissement du méta-modèle 3M. Chaque enrichissement définit ses propres spécialisations de Descriptor. Les terminologies incluses dans ce modèle sont mises en œuvre sous forme d'ontologies OWL légères. Le passage d'une représentation ontologique à une terminologique est basé sur un processus de réification. De cette façon, les ontologies formelles sont « dégradées » pour s'adapter à ce modèle. Cette méthode permet de conserver les données d'origine d'une ontologie et d'enrichir le serveur avec des ressources terminologiques de santé. L'approche conceptuelle utilisée dans le modèle permet l'intégration de toute langue tout en maintenant les relations entre les concepts.

Méthode d'intégration

Le modèle physique de données CISMeF se compose ainsi de deux parties : le « modèle » pour définir un modèle conceptuel de données et d'autre part l'instance du modèle pour stocker les données elle-mêmes (voir la figure A.1 en annexe). La partie « modèle » se compose de quatre tables préfixées de « `MODEL` », en bas de la figure. Les entités sont stockées dans la table `TB_MODEL`, les attributs dans la table `TB_MODEL_DATATYPE_PROPERTY` et les associations entre entités dans la table `TB_MODEL_DATATYPE_PROPERTY`. La table `TB_MODEL_INHERITANCE` permet de gérer les relations d'héritage. Les cinq tables `TB_OBJECT`, `TB_OBJECT_PROPERTY`, `TB_DATATYPE_PROPERTY`, `TB_HIERARCHY` et `TB_INDEXING` permettent de gérer les objets concrets ainsi que leurs attributs et relations. Pour intégrer des terminologies dans le système d'information CISMeF, trois étapes sont nécessaires : (i) concevoir un modèle générique de terminologie dans lequel chaque terminologie peut être intégré, (ii) concevoir un processus capable d'intégrer des terminologies dans la base de données implémentant le modèle et (iii)) construire et intégrer les relations sémantiques intra et inter-terminologie. Un modèle générique est disponible afin de tenir compte de toutes les terminologies dans une structure globale. Ensuite, un modèle de chaque terminologie est conçu comme une spécialisation du méta-modèle. Avec les modèles spécifiques, le travail consiste à développer un analyseur pour chaque terminologie : l'entrée est la donnée d'origine (ou les données originales normalisées) et la sortie est un fichier OWL dans le cas d'une ontologie ou un fichier RDF/XML dans le cas d'une terminologie, cas appliqué ici pour l'intégration des bases NCBI Gene, Uniprot Swissprot KB et OMIM. Comme les données peuvent être dans différents formats et structures, dans certains cas, des processus supplémentaires doivent être effectués (bases de données temporaires, fichiers, etc.).

La phase finale est l'intégration des fichiers OWL ou RDF/XML dans le modèle physique. Un parseur générique a été développé pour insérer directement chaque terminologie dans la base de données, capable de reconnaître les classes descriptrices, les définitions, les synonymes, les relations afin de l'insérer très facilement dans la base de données.

Implémentation du modèle de données omiques

Préalablement à l'intégration de données dans le modèle décrit, Le modèle de données omiques a dû être implémenté au sein du méta modèle CISMeF, en lien avec le modèle de données RAVEL. J'ai pour cela défini le modèle conceptuel des données omiques dans la partie « modèle générique » du méta-modèle 3M.

4.1.3 Données de référence

Le modèle de données développé s'appuie ainsi sur l'intégration de données de références externes, concernant les gènes et protéines ainsi que les pathologies associées afin de faciliter l'interrogation des données et la recherche d'information. L'intégration de ces données dans le système d'information du D2IM du CHU de Rouen fait donc appel aux étapes décrites précédemment : la création d'un modèle conceptuel des données, la création d'un parseur des données sources et enfin l'intégration en base de données.

NCBI Gene

Le NCBI met à disposition l'ensemble des données de la base Gene sur son serveur FTP, mises à jour quotidiennement, sous la forme d'un fichier binaire `.xgs`.

La première étape du traitement (voir FIGURE 4.1) consiste à convertir ce fichier en un fichier XML exploitable, contenant uniquement les gènes identifiés chez *Homo sapiens*. Pour cela, il est traité avec un script bash, fourni par le NCBI. Le fichier XML obtenu contient l'intégralité des gènes et microARN identifiés chez l'Homme et représente environ 9 Go.

La seconde étape consiste à parser ce fichier XML afin d'insérer les données dans la base de données grâce au parseur Java P1. Pour cela, la classe `ParseurGene.java` a été développée. Cette classe extrait les informations correspondantes au modèle de données établi en analysant les balises du document XML et réalise l'insertion en base de données.

Afin de suivre les mises à jour effectuées chaque jour par le NCBI, le parseur P1 a été modifié afin qu'il soit exécutable en ligne de commande. L'ensemble des processus (téléchargement du fichier source, conversion et analyse) est ainsi exécuté grâce à un script bash, dont le lancement est programmé quotidiennement.

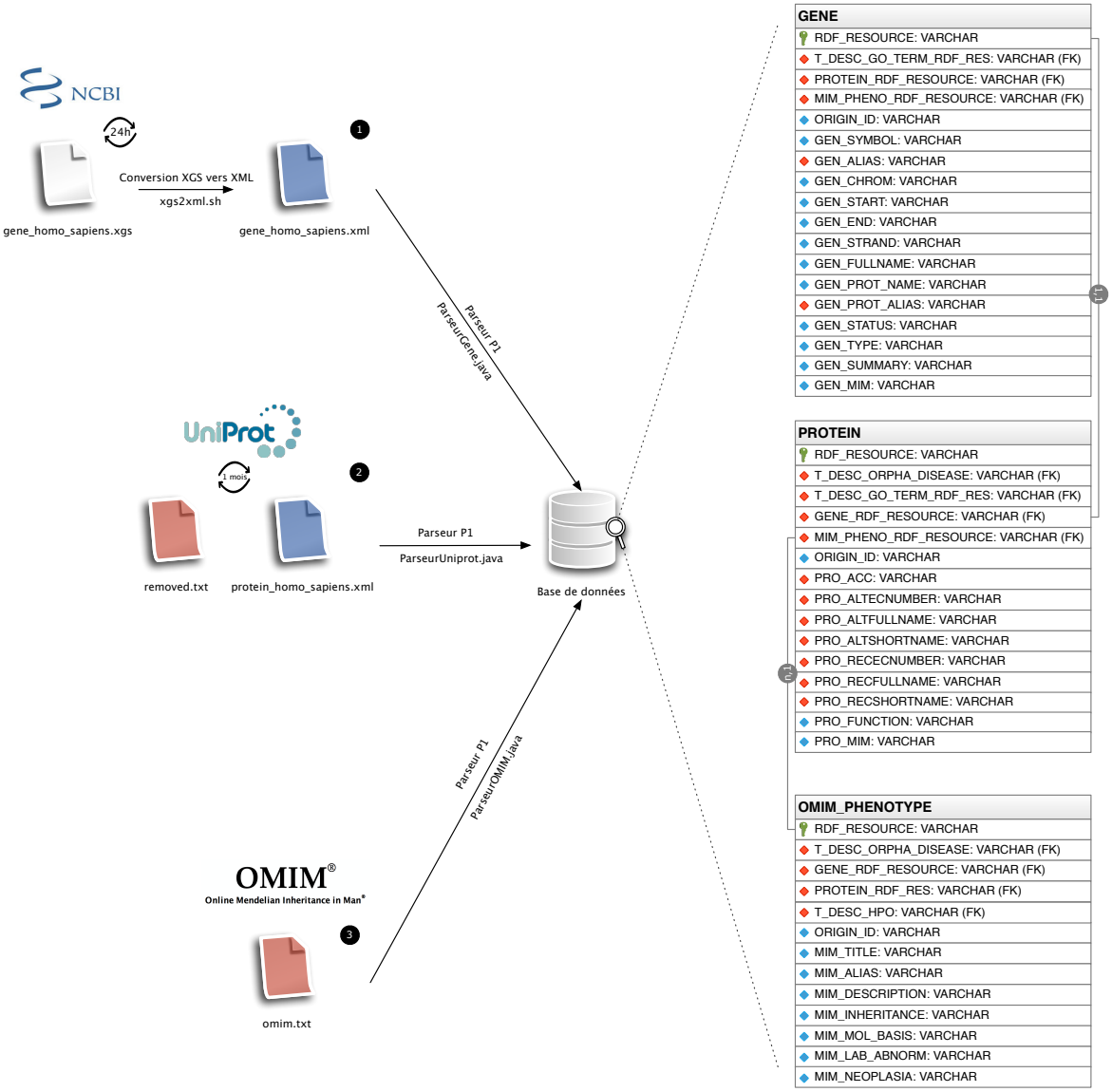

FIGURE 4.1 – Processus d'intégration des données externes.

UniprotKB

Uniprot met à disposition l'intégralité de sa base UniprotKB/Swissprot sous la forme d'un fichier XML d'environ 800 Mo et les données obsolètes sont mises à disposition par un fichier texte. Leur traitement est décrit sur la FIGURE 4.1.

L'analyse des deux fichiers source, fichier de données XML et fichier texte contenant les données obsolètes, est réalisée par une classe Java `ParseurUniprot` du parseur P1 qui extrait les données sélectionnées dans le modèle de données et les insère en base de données. Les données déclarées obsolètes sont supprimées.

Uniprot met à jour le fichier de données sources de manière irrégulière, environ une fois par mois. Ainsi la mise à jour des données ne peut pas être ici automatisée et doit être réalisée manuellement.

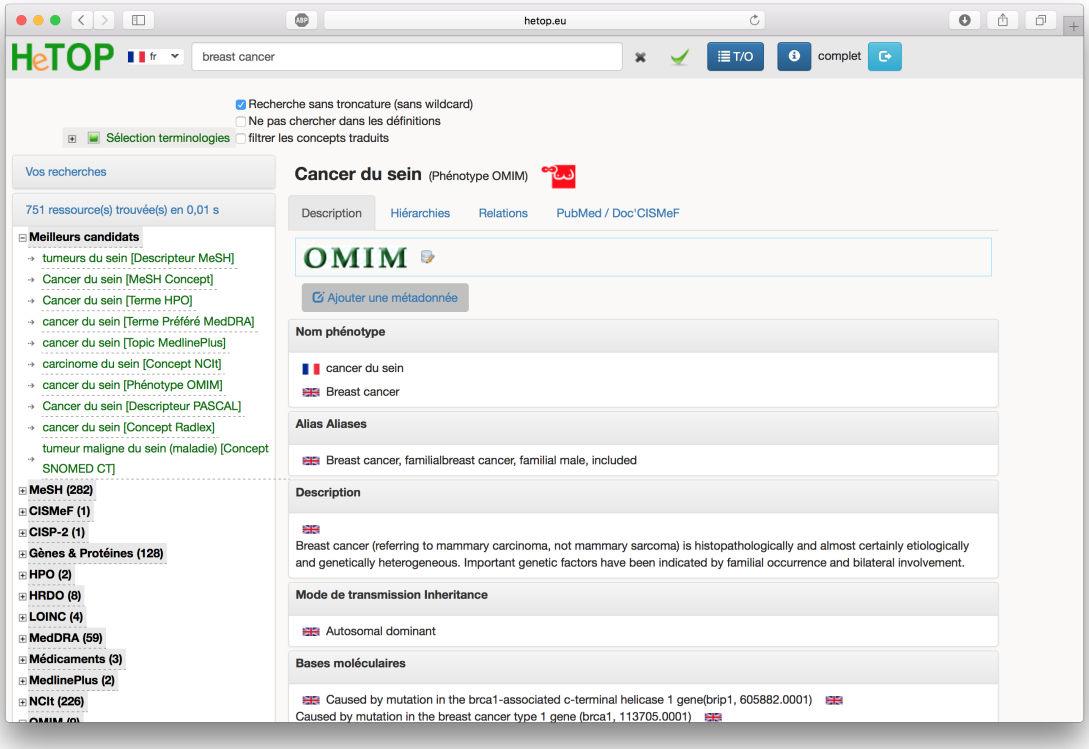

FIGURE 4.2 – Capture d'écran de la page de description d'une pathologie OMIM : le cancer du sein.

OMIM

OMIM met à disposition ses données sous la forme d'un fichier texte non structuré ou d'une API. L'accès par l'API étant limité à dix enregistrements, le fichier texte non structuré a dû être utilisé.

Les données concernant les gènes ayant déjà été importées de la base Gene du NCBI, seules les données concernant les pathologies ont été traitées. Certaines entrées combinant description d'un gène et d'une pathologie ont également été traitées.

L'analyse des données est réalisée par la classe Java `ParseurOMIM` (voir FIGURE 4.1). Les données étant non structurées, le traitement des données repose principalement sur des expressions régulières ciblant les informations correspondant au modèle de données. Les données extraites sont ensuite insérées en base de données.

La fréquence des mises à jour n'est pas communiquée par OMIM, elles doivent donc être lancées manuellement.

Les données intégrées ont été mises à disposition sur le portail HeTOP (voir FIGURE 4.2) en libre accès. On peut ainsi retrouver grâce au moteur de recherche les informations concernant les gènes, protéines et pathologies OMIM intégrées ainsi que les données associées qui étaient déjà présentes sur le portail (annotations GO, maladies Orphanet et description des phénotypes grâce à la terminologie HPO). La figure 4.2

montre les informations extraites des données d'OMIM et disponibles dans la base de données : nom du phénotype, alias de la maladie, description, mode de transcription, base moléculaire et type de la tumeur dans le cas d'un cancer. L'onglet Relations visible en haut de la page répertorie les différentes relations avec les autres terminologies et ontologies disponibles. Ici, il s'agit du gène et de la protéine liés ainsi que des maladies Orphanet correspondantes.

4.1.4 Données expérimentales

L'intégration de données expérimentales lève une nouvelle problématique liée à la structuration des données. D'une part, les fichiers de données décrivant une étude sont standardisés et adoptent le format IDF, correspondant à un fichier texte tabulé structuré (voir FIGURE 4.3b). Les différents champs (nom de l'étude, nom des responsables, objet de l'étude) sont constants. D'autre part, les résultats expérimentaux sont présentés le plus souvent dans des fichiers texte tabulés (voir FIGURE 4.3a). Un fichier correspond à un échantillon, conformément au niveau d'interprétation 3 choisi. Ces fichiers sont peu structurés et inconstants d'un type d'étude à l'autre, en fonction de la nature de l'analyse, de la technique utilisée voire du laboratoire impliqué.

L'intégration de ces données a donc nécessité l'élaboration d'un format intermédiaire. Pour cela, j'ai choisi de concevoir un document XML Schema Definition (XSD) permettant de définir la structure et le type de contenu d'un document XML et de vérifier la validité de ce document. Ainsi le schéma XSD conçu définit de manière exhaustive les différents champs nécessaires à la description de toute étude omique expérimentale, quelle que soit la technique utilisée ou la nature de l'expérience (voir FIGURE 4.3c). Il permet de réunir en un fichier unique et standard les informations concernant l'étude issues du fichier IDF ainsi que les résultats expérimentaux issues des multiples fichiers de résultats.

Le processus d'intégration des données se déroule en deux étapes. Tout d'abord, les différents fichiers comportant les informations annexes à l'expérience et les résultats expérimentaux sont traités en Python afin de les convertir en un fichier XML unique répondant au schéma XSD élaboré (voir FIGURE 4.3d). Ensuite, ce fichier XML est traité en Java, via une classe dédiée dans le parseur P1, afin d'en extraire les informations et d'insérer les données dans la base de données. Le traitement en Java peut inclure certaines étapes d'annotation, afin par exemple de déterminer les gènes présents sur un segment génomique dans le cadre de l'analyse de CNV. Ces étapes sont décrites sur la FIGURE 4.5.

La conversion en XML des données expérimentales nécessite de développer un script Python (ou tout autre langage de programmation adapté au traitement de fichiers) spécifique à chaque format de fichier de résultats expérimentaux. Cependant un module Python dédié au traitement des fichiers IDF a été développé, les formats de ces fi-

```
barcode antibody name   gene name       protein expression value
DM_PAT_1098     14-3-3_epsilon-M-C      YWHAE   -.020482195
DM_PAT_1098     4E-BP1-R-V      EIF4EBP1        -2.230901393
DM_PAT_1098     4E-BP1_pS65-R-V EIF4EBP1        -.857013273
DM_PAT_1098     4E-BP1_pT37-R-V EIF4EBP1        .568021218
DM_PAT_1098     4E-BP1_pT70-R-C EIF4EBP1        -1.042769877
DM_PAT_1098     53BP1-R-C       TP53BP1 -.4017518
DM_PAT_1098     ACC1-R-C        ACACA   2.468771596
DM_PAT_1098     ACC_pS79-R-V    ACACAACACB      -.904404774
DM_PAT_1098     AIB1-M-V        NCOA3   -.00820980499999999
DM_PAT_1098     AMPK_alpha-R-C  PRKAA1  .757769158
DM_PAT_1098     AMPK_pT172-R-V  PRKAA1  1.426208884
DM_PAT_1098     AR-R-V  AR      -1.456589894
DM_PAT_1098     ARID1A-M-V      ARID1A  .306226933
DM_PAT_1098     ATM-R-C ATM     -2.613545584
DM_PAT_1098     Akt-R-V AKT1AKT2 AKT3   1.569291165
DM_PAT_1098     Akt_pS473-R-V   AKT1AKT2 AKT3   -.757387354
DM_PAT_1098     Akt_pT308-R-V   AKT1AKT2 AKT3   -.770316686
DM_PAT_1098     Annexin_I-R-V   ANXA1   -.942110595
DM_PAT_1098     B-Raf-M-NA      BRAF    1.538239405
DM_PAT_1098     Bak-R-C BAK1    -2.823334273
DM_PAT_1098     Bax-R-V BAX     -.668640959
DM_PAT_1098     Bcl-2-R-NA      BCL2    -2.668710509
DM_PAT_1098     Bcl-X-R-C       BCL2L1  -1.904491073
DM_PAT_1098     Bcl-xL-R-V      BCL2L1  -1.490967658
DM_PAT_1098     Beclin-G-V      BECN1   -1.294727351
DM_PAT_1098     Bid-R-C BID     -2.841195061
DM_PAT_1098     Bim-R-V BCL2L11 -.794844451
DM_PAT_1098     C-Raf-R-V       RAF1    .331993678
DM_PAT_1098     C-Raf_pS338-R-C RAF1    -1.798225777
DM_PAT_1098     CD20-R-C        CD20    -1.936650806
DM_PAT_1098     CD31-M-V        PECAM1  -1.268609027
DM_PAT_1098     CD49b-M-V       CD49    .145377507
DM_PAT_1098     CDK1-R-V        CDC2    -.919522635
DM_PAT_1098     COX-2-R-C       PTGS2   -2.111512029
DM_PAT_1098     Caspase-3_active-R-C    CASP3   -1.75977991
DM_PAT_1098     Caspase-7_cleavedD198-R-C       CASP7   -3.599636766
DM_PAT_1098     Caspase-8-M-C   CASP8   .273824069
DM_PAT_1098     Caspase-9_cleavedD330-R-C       CASP9   -1.474111013
DM_PAT_1098     Caveolin-1-R-V  CAV1    4.213981068
DM_PAT_1098     Chk1-R-C        CHEK1   .695998622
DM_PAT_1098     Chk1_pS345-R-C  CHEK1   -2.158530813
DM_PAT_1098     Chk2-M-C        CHEK2   -.125000606
```

(a) Fichier de résultats non traité.

```
Investigation Title     Protein Expression for TCGA glioblastoma multiforme
Protein Array (RPPA)
Experimental Design     is_expressed_design
Experimental Design Term Source REF     MGED Ontology   MGED Ontology
Experimental Factor Type        disease_state_design
Experimental Factor Type Term Source REF        MGED Ontology

Person Last Name        Mills
Person First Name       Gordon
Person Mid Initials     B
Person Email    gmills@mdanderson.org
Person Phone    713-563-4200
Person Address  "1400 Pressler Street, Unit Number: 1410, Houston, TX 77030"

Person Affiliation      The University of Texas MD Anderson Cancer Center

Person Roles    PI

Date of Experiment      2011-10-05

Public Release Date     2011-11-16

Experiment Description  Protein Expression for TCGA GBM samples using MDA Re


Protocol Name   mdanderson.org:protein_extraction:MDA_RPPA_Core:01      mdan
01      mdanderson.org:serial_dilution:MDA_RPPA_Core:01 mdanderson.org:slide
mdanderson.org:antibody_probe:MDA_RPPA_Core:01  mdanderson.org:slide_scan:MD
mdanderson.org:supercurve:MDA_RPPA_Core:01      mdanderson.org:protein_norma
Protocol Type   protein_extraction      protein_denature        serial_dilut
slide_scan      supercurve      protein_normalization
Protocol Term Source REF        NCI EVS NCI EVS NCI EVS NCI EVS NCI EVS NCI
Protocol Description    protein extracted from shipped portions sent by BCR.
fold serial dilution using dilution bufer       print (Aushon Biosystems 247
Labs Oncyte Avid nitrocellulose film slides)    wash antibody over slide (Da
(Cannon Scan 9000F)     ran SuperCureve R-package on data       Loading corr
Protocol Parameters

SDRF Files      mdanderson.org_GBM.MDA_RPPA_Core.sdrf.txt
Term Source Name        MGED Ontology   NCI EVS NCBI Taxonomy
Term Source File        http://mged.sourceforge.net/ontologies/MGEDontology.
www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/
Term Source Version     1.3.1.1 2010-09 2010-09
```

(b) Fichier IDF.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<!-- definition of complex elements -->
        <xs:element name="omi_lab">
          <xs:complexType>
                  <xs:element ref="has_country"/>
                  <xs:element ref="has_city"/>
                  <xs:element ref="omi_lab_ref"/>
                  <xs:element ref="omi_lab_name"/>
                  <xs:element ref="omi_lab_team"/>
                  <xs:element ref="omi_lab_address"/>
                  <xs:element ref="omi_lab_mail"/>
                  <xs:element ref="omi_lab_phone"/>
          </xs:complexType>
        </xs:element>
        <xs:element name="omi_study">
          <xs:complexType>
                  <xs:element ref="omi_lab_rdf"/>
                  <xs:element ref="omi_study_has_unit"/>
                  <xs:element ref="omi_std_head_firstname"/>
                  <xs:element ref="omi_std_head_lastname"/>
                  <xs:element ref="omi_std_title"/>
                  <xs:element ref="omi_std_type"/>
                  <xs:element ref="omi_std_protocol"/>
                  <xs:element ref="omi_std_equipement"/>
                  <xs:element ref="omi_std_desc"/>
                  <xs:element ref="omi_std_sample"/>
                  <xs:element ref="omi_std_date"/>
                  <xs:element ref="omi_std_source"/>
          </xs:complexType>
        </xs:element>
        <xs:element name="omi_submitter">
      <xs:complexType>
          <xs:element ref="omi_sub_firstname"/>
          <xs:element ref="omi_sub_lastname"/>
          <xs:element ref="omi_sub_team"/>
          <xs:element ref="omi_sub_mail"/>
          <xs:element ref="omi_sub_phone"/>
      </xs:complexType>
    </xs:element>
        <xs:element name="omi_units">
          <xs:complexType>
                  <xs:element ref="omi_unit"/>
```

(c) Schéma XSD.

```
<omic_study_track xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:
    <omi_lab>
        <omi_lab_address>
            1400 Pressler Street, Unit Number: 1410, Houston, TX 77030
        </omi_lab_address>
        <omi_lab_name>
            The University of Texas MD Anderson Cancer Center
        </omi_lab_name>
    </omi_lab>
    <omi_study>
        <omi_std_title>
            Protein Expression for TCGA glioblastoma multiforme (GBM) sample
(RPPA)
        </omi_std_title>
        <omi_std_date>
            2011-10-05
        </omi_std_date>
        <omi_std_desc>
            Protein Expression for TCGA GBM samples using MDA Reverse Phase
        </omi_std_desc>
        [...]
    </omi_study>
    <omi_units />
    <omi_submitter>
        <omi_sub_lastname>
            Mills
        </omi_sub_lastname>
        <omi_sub_firstname>
            Gordon
        </omi_sub_firstname>
        [...]
    </omi_submitter>
    <omi_gene>
        <omi_quantif_has_gene>
            YWHAE
        </omi_quantif_has_gene>
        <omi_quantif>
            <omi_quantif_has_pat>
                DM_PAT_1098
            </omi_quantif_has_pat>
            <omi_quantif_val>
                -0.020482195
            </omi_quantif_val>
```

(d) Exemple de sortie XML.

FIGURE 4.4 – Traitement des données omiques expérimentales.

(a) Extrait d'un fichier texte tabulé de résultats d'analyse d'expression de protéines avant traitement pour un échantillon.

(b) Extrait du fichier IDF décrivant les paramètres de l'étude.

(c) Extrait du schéma XSD défini pour la conversion des données en XML.

(d) Extrait du fichier XML obtenu après traitement, comportant à la fois les données de l'étude issues du fichier IDF et les résultats expérimentaux.

chiers étant constants d'une étude à l'autre. L'intégration d'un nouveau type d'étude omique ne demande ainsi que le développement d'un script de traitement des fichiers de résultats expérimentaux.

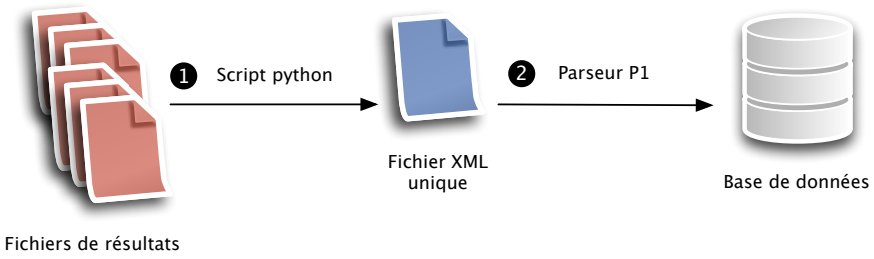

FIGURE 4.5 – Intégration des données omiques expérimentales.

4.1.5 Lien avec les données cliniques

Préalablement à l'intégration des données, il a été nécessaire d'établir une méthode afin d'associer les patients présents dans la base de données RAVEL et les fichiers de résultats omiques expérimentaux à disposition. En effet, les données omiques récoltées n'étaient pas liées à des données cliniques disponibles. Il a donc fallu relier les données cliniques déjà présentes dans RAVEL avec les données omiques réunies. Pour cela, la classification CIM10 a été utilisée. Chaque patient dont les données sont intégrées à RAVEL est associé à plusieurs codes CIM10, en fonction des pathologies diagnostiquées. À partir de ces codes, les patients dont les pathologies correspondaient aux pathologies pour lesquelles des données omiques étaient disponibles ont été sélectionnés. Ainsi, la cohérence entre les données de RAVEL et les données omiques issues du portail TGCA ou fournies par les laboratoires contactés a été préservée.

Le processus d'intégration a été utilisé pour intégrer avec succès 7 études omiques dont 5 études issues du projet « Glioblastoma » de la base de données TGCA, et 2 études dont les données ont été obtenues par collaboration. Ces études portent sur (i) l'analyse de méthylation de l'ADN, (ii) analyse CGH, (iii) analyse de CNV, (iv) analyse de l'expression des gènes, (v) analyse de l'expression de protéines, (vi) analyse des miRNA et (vii) l'analyse d'un exome. Au total les données de 527 patients ont été intégrées dans notre base de données cliniques et omiques.

4.2 Recherche d'information clinomique

4.2.1 Moteur de recherche CISMeF

Par rapport à l'existant, le moteur de recherche utilisé dans RAVEL a été conçu à la fois comme le plus générique possible et avec une forte contrainte en terme de temps

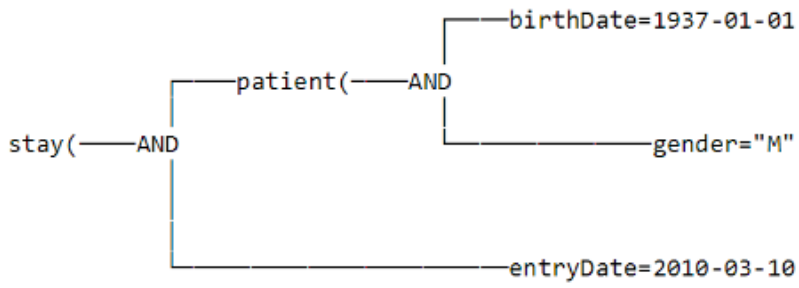

FIGURE 4.6 – Représentation en arbre de la requête `stay(patient(birthDate=1937-01-01 AND gender="M") AND entryDate=2010-03-10)`.

de réponse. Cet outil s'appuie sur les travaux précédents de l'équipe CISMeF sur la RI documentaire, dans le cadre desquels a été créé un moteur sémantique pour retrouver un type d'objet unique : les ressources web. Ce moteur de recherche permet la recherche d'information sur des données structurées, essentiellement numériques, mais aussi parfois symboliques (comme le genre), mais aussi sur des données non structurées (issues essentiellement des différents comptes-rendus d'un DPI). Ce moteur est générique, car il permet de rechercher n'importe quelles données stockées dans la base de données.

Ce moteur de recherche se base sur un langage de requêtes dédié qui présente trois caractéristiques importantes : (i) il est orienté objet, (ii) il est flexible et évolutif et (iii) il a des capacités d'interrogation complète, c'est-à-dire que toutes les données incluses dans la base de données sont interrogeables. Le moteur de recherche a été conçu pour être générique, rapide, multilingue et aligné avec de multiples terminologies. Il utilise un langage d'interrogation spécifique visant à faciliter la recherche d'information. Le langage est composé d'unités syntaxiques, respectant la syntaxe suivante :

`ENTITÉ(CLAUSE_CONTRAINTES)`

où :

— `ENTITÉ` correspond à une entité du modèle conceptuel

— `CLAUSE_CONTRAINTES` correspond aux contraintes appliquées à cette entité, construite en utilisant des attributs ou des objets liés (voir TABLEAU 4.1).

Plus de détails sont disponibles sur le langage d'interrogation dans [LELONG et al., 2016]. Dans le modèle actuel, trois `ENTITÉS` principales sont modélisées à trois niveaux : le niveau du patient, le niveau du séjour, et le niveau le plus bas (comme l'analyse biologique ou l'analyse omique). Par exemple, les requêtes `patient()` et `medicalUnit()` retournent respectivement tous les patients et toutes les unités médicales contenues dans la base données. La clause de contrainte `CLAUSE_CONTRAINTES` permet d'appliquer des contraintes à l'entité spécifiée. C'est une expression booléenne, ainsi les opérateurs booléens `AND, OR, NOT` et les parenthèses sont utilisées pour construire des liens logiques

entre contraintes uniques. Cette clause de contrainte peut être construite en utilisant les attributs de l'entité spécifiée. Par exemple, la requête suivante `patient(birthDate= 1937-01-01 AND gender='M')` utilise deux attributs `birthDate` et `gender` de l'entité `patient` et retournera tous les patients masculins, nés le 01/01/1937. Les opérateurs booléens, parenthèses et comparateurs sont définis explicitement dans la grammaire du langage alors que les entités sont déduites automatiquement par auto-complétion à partir du modèle de données omiques. Le moteur de recherche permet d'interpréter les requêtes pour extraire les données correspondantes de la base de données. Le processus d'interprétation contient trois étapes : (i) le parsing de la requête, (ii) sa représentation sous la forme d'un arbre (voir FIGURE 4.6) et (iii) la construction de la requête SQL correspondant à l'arbre généré, le modèle de données étant intégré dans une base de données relationnelle. Différentes données peuvent être extraites : (i) données symboliques (absence, présence), (ii) données numériques (avec les opérateurs >, < et =) et (iii) données chronologiques.

Requête en langage naturel	**Traduction dans le langage d'interrogation**

Les patients de l'étude 12 ayant une expression de HRNR supérieure à 3	`patient(`
	`study(id="OMI_STUDY_12")`
	`AND quantification(gene(geneSymbol="HRNR")`
	`AND numericValue > 3)`

Patients ayant des variations faux-sens sur HOMER1 et un taux de glucose sanguin supérieur à 1,1g/L	`patient(`
	`study(id="OMI_STUDY_1")`
	`AND variant(gene(geneSymbol="HOMER1")`
	`AND variantCategory="Missense" )`
	`AND bioTest(`
	`bioResultEXECode(label="Glucose")`
	`AND numericValue > 1.1))`

Tous les segments génomiques délétés dans l'étude 10	`quantification(interpretation="deletion"` `AND study(id="OMI_STUDY_10"))`

Tous les variants faux sens sur le gène HRNR sans l'étude 1	`variant(study(id="OMI_STUDY_1")`
	`AND gene(geneSymbol="HRNR")`
	`AND variantCategory="Missense")`

TABLEAU 4.1 – Exemples de requêtes omiques.

4.2.2 Requête dans les données omiques

Description

Le moteur de recherche a été adapté pour permettre aux utilisateurs d'interroger à la fois données cliniques et données omiques (variant, gène, protéine ou segment génomique). Des mots-clés ont été définis pour interroger chaque entité du modèle omique conceptuel et élaborer des contraintes. Le temps de réponse moyen pour un patient est inférieur à deux secondes, ce qui est considéré comme satisfaisant pour un clinicien ou un chercheur. Pour n patients, le temps de réponse moyen est inférieur à dix secondes. Il est possible de réaliser la RI simultanément sur les données cliniques et omiques dans la même requête. Par exemple, les patients qui ont des variations faux-sens sur le gène HOMER1 et un taux de glucose sanguin supérieur à 1,1g/L peuvent être extraits. Chaque entité du modèle de données omiques conceptuel est interrogeable. Les requêtes sont réalisables à plusieurs échelles : au niveau du patient, du séjour ou de l'étude expérimentales. Les variants et régions génomiques peuvent également être extraits.

Interrogation sémantique des données

Le système fourni permet ainsi d'interroger les données cliniques et omiques dans un ou plusieurs DPI. Plusieurs types de données omiques peuvent être récupérés : données d'expression (gènes, protéines, miARN), variantes génomiques, variantes de numéros de copies, hybridation génomique comparative et données de méthylation d'ADN. Les terminologies et les ontologies telles que SNOMED CT, MeSH, CIM10 utilisées comme références dans l'interface utilisateur graphique peuvent être utilisées pour créer des requêtes. Par exemple, tous les patients ayant des variantes génomiques sur des gènes annotés avec le terme Gene Ontology `GO:0042246` *Récupération tissulaire* peuvent être récupérés avec la requête `patient(variant(gene(GOterm(label="tissue regeneration"))))`.

Visualisation des résultats

Les données omiques expérimentales sont consultables dans un onglet réservé du prototype RAVEL qui regroupe l'intégralité des différents types de données (voir Figure 4.7). Chaque analyse est représentée dans un cadre propre, les résultats apparaissant dans un tableau au sein de ce cadre. Si l'analyse considérée est une analyse quantitative (analyse d'expression, de méthylation...), les informations affichées sont : (i) l'entité analysée (gène, protéine, segment analysé), (ii) la valeur obtenue, normalisée, (iii) l'interprétation de cette valeur (« délétion », « sous-expression »...), (iv) un commentaire (facultatif), (v) la date de la mesure et (vi) un lien vers l'étude afférente.

Si l'analyse considérée concerne un variant (SNP ou indel), les informations affichées sont : (i) le nom du variant, (ii) le nom du variant en accord avec la nomenclature

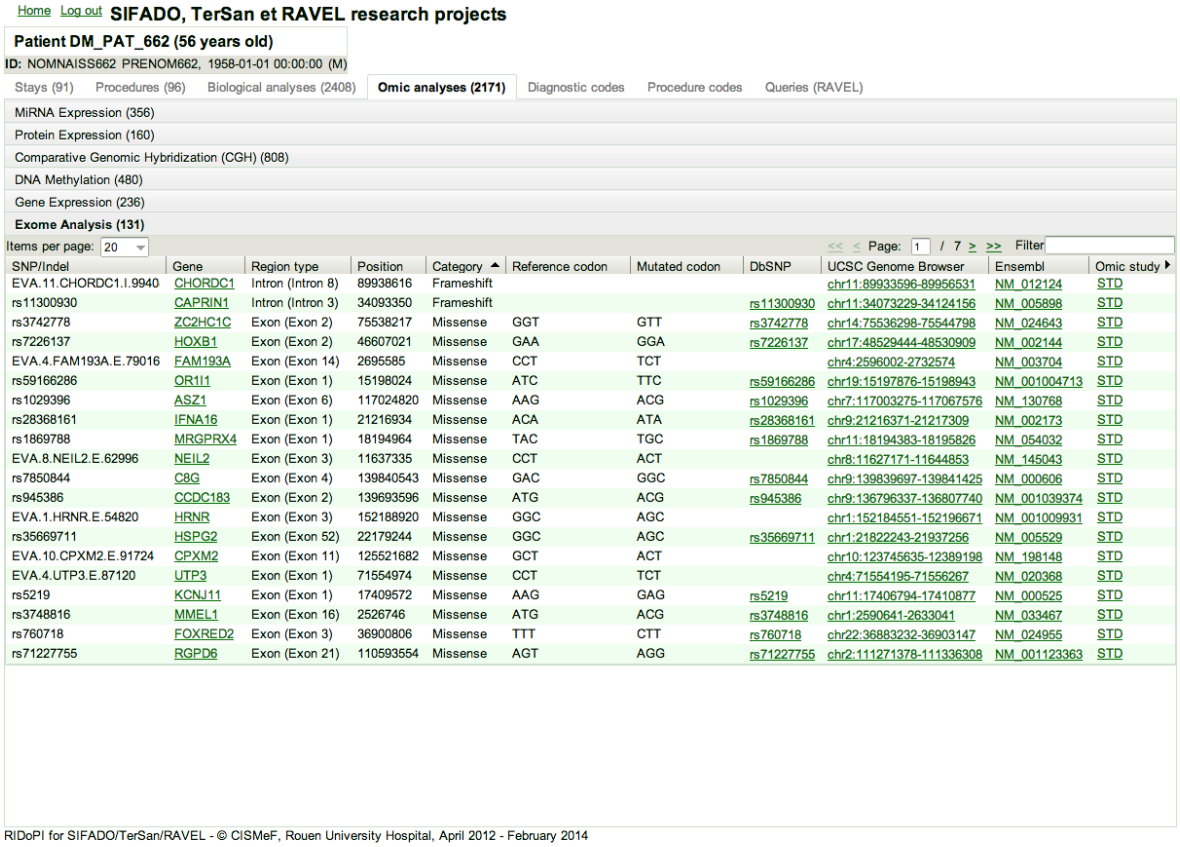

FIGURE 4.7 – Capture d'écran du prototype RAVEL.

HGVS si disponible, (iii) le gène, (iv) le type de région, (v) la position du variant, (vi) la catégorie du variant, (vii) le codon de référence, (viii) le codon muté, (ix) un lien vers la base de données dbSNP répertoriant les SNP présents chez différentes espèces dont l'Homme [SHERRY et al., 2001], (x) un lien vers le Genome Browser de l'UCSC permettant d'explorer le génome, (xi) un lien vers la base de données de génomes eucaryotes Ensembl, (xii) la date de l'analyse, et (xiii) un lien vers l'étude afférente.

La FIGURE 4.7 présente la vue principale des données omiques au sein de l'application RAVEL développée par l'équipe CISMeF. Les données sont présentées au sein d'un composant en « accordéon », chaque onglet étant dédié à un type d'analyse omique. Ici on peut observer chez ce patient ces résultats de (i) analyse miARN, (ii) expression protéique, (iii) CGH, (iv) analyse de méthylation de l'ADN, (v) expression génique et (vi) analyse d'exome.

Le détail de chaque entité analysée (gène, protéine, miARN, segment génomique) peut être consulté dans l'interface. Chacune des entités bénéficie des données de références agrégées depuis NCBI Gene et Uniprot KB. La FIGURE 4.8 présente les informations concernant le gène SAGE1 issues de la base de données NCBI Gene. Toutes les informations de référence importées depuis NCBI Gene sont consultables directement au sein de l'application. Ces informations incluent les informations générales concer-

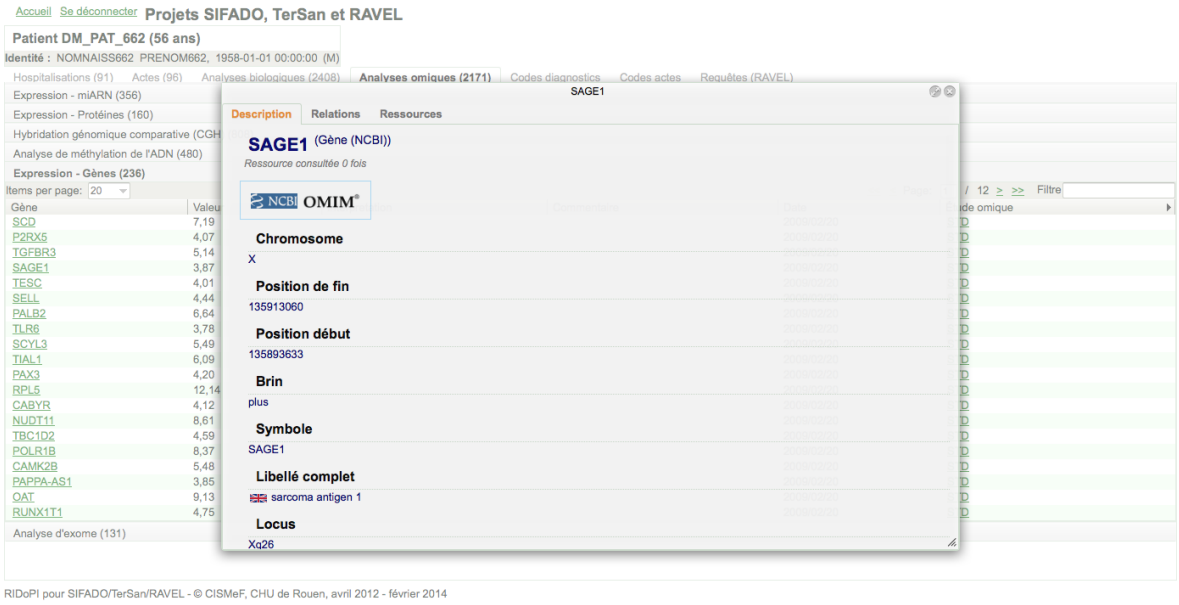

FIGURE 4.8 – Détail concernant le gène SAGE1 depuis l'interface de visualisation des données omiques.

nant le gène (nom, alias, position, fonction) ainsi que les relations (annotations GO, protéines associées, pathologies associées) et ressources liées.

Les données concernant les études omiques sont consultables également depuis le prototype RIDoPI dans une fenêtre dédiée. Ces informations incluent notamment (i) la date de l'étude, (ii) le titre de l'étude, (iii) sa description, (iv) l'équipement utilisé, (v) le protocole utilisé, (vi) la source des données, (vii) le type de l'étude (expression de gènes, analyse de variants...) et (viii) si applicable, la version d'assemblage du génome utilisée.

L'utilisateur a ainsi la possibilité de consulter la source des données qu'il visualise, notamment de s'informer sur le moyen d'obtention des échantillons analysés ou encore le protocole utilisé pour le traitement des données. Il peut également de cette façon obtenir les données brutes associées aux données disponibles, afin par exemple de les réutiliser ou effectuer un nouveau traitement de ces données brutes.

De plus, une vue globale des données de l'étude est possible à travers différents graphes. La FIGURE 4.9 montre les informations concernant l'étude « miRNA Analysis of TCGA GBM Samples » portant sur 54 patients atteints d'un glioblastome dont les résultats omiques ont été publiés par le TCGA. Toutes les informations apparaissant dans le fichier idf fourni (voir FIGURE 4.3b) sont stockées en base de données et sont consultables directement au sein de l'application, incluant notamment le titre de l'étude, sa description et le protocole utilisé. Ces données regroupent le nombre de patients, les caractéristiques démographiques de la cohorte (âge, genre, origine ethnique...). Elles sont représentées à partir d'histogrammes et de diagrammes circulaires, afin de pouvoir rapidement appréhender les caractéristiques de l'étude.

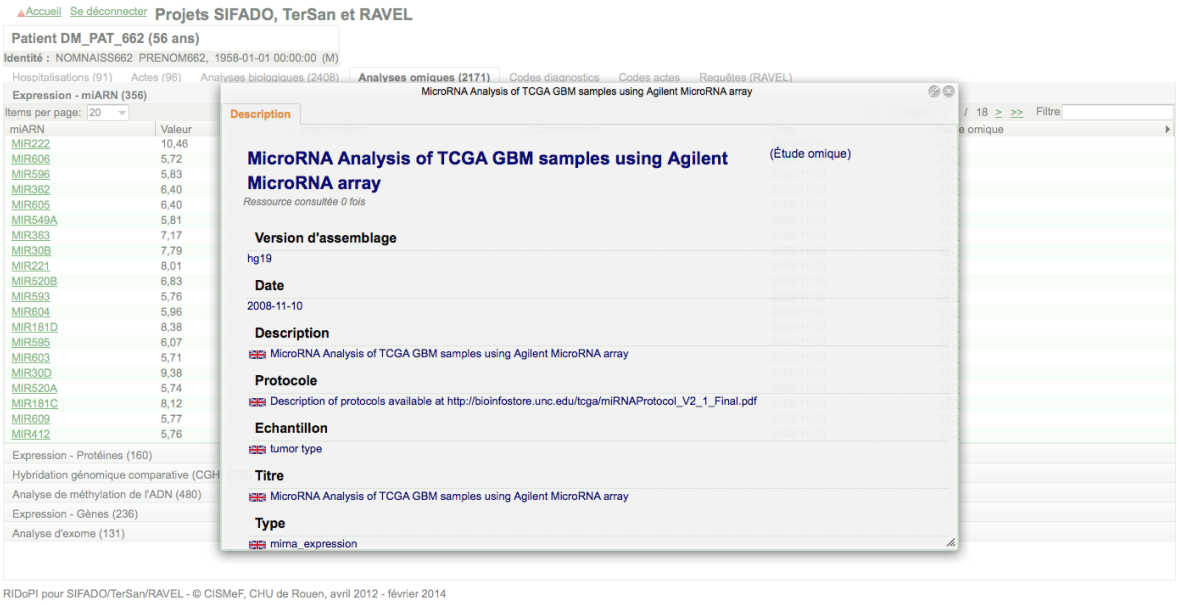

FIGURE 4.9 – Détail concernant l'étude « miRNA Analysis of TCGA GBM Samples » depuis l'interface de visualisation des données omiques.

4.3 Résultats

4.3.1 Comparatif face aux solutions existantes : i2b2 et tranSMART

i2b2

Afin d'étudier les solutions existantes dédiées à la gestion et l'exploitation des données issues des DPI, j'ai procédé à l'installation des systèmes i2b2 et tranSMART. D'une part, j'ai étudié leur architecture serveur et d'autre part les logiciels clients ainsi que les fonctionnalités proposées.

i2b2 est une plateforme libre développée aux États-Unis et dédiée à la recherche translationnelle. Cette plateforme est implémentée dans plusieurs pays ([TAKAI-IGARASHI et al., 2011] et [GANSLANDT et al., 2011]) et est devenue un standard *de facto*. i2b2 stocke les données dans un entrepôt de données Oracle centré sur l'exploitation de données structurées. L'architecture fonctionnelle de cet outil écrit en Java propose une organisation en modules (*hive*), chaque module assurant une fonction spécifique au sein de l'applicatif. Bien que cette modularité facilite l'utilisation des données des DPI pour les chercheurs, elle complexifie le déploiement et la maintenabilité du système. Ces modules incluent (i) des outils de traitement automatique de la langue pour l'indexation et l'extraction de concepts basés sur des ontologies médicales, (ii) des outils de recherche d'information pour sélectionner des patients sur des données codées et structurées, (iii) des outils graphiques pour réaliser des analyses statistiques sur les résultats de la sélection [MURPHY et al., 2006].

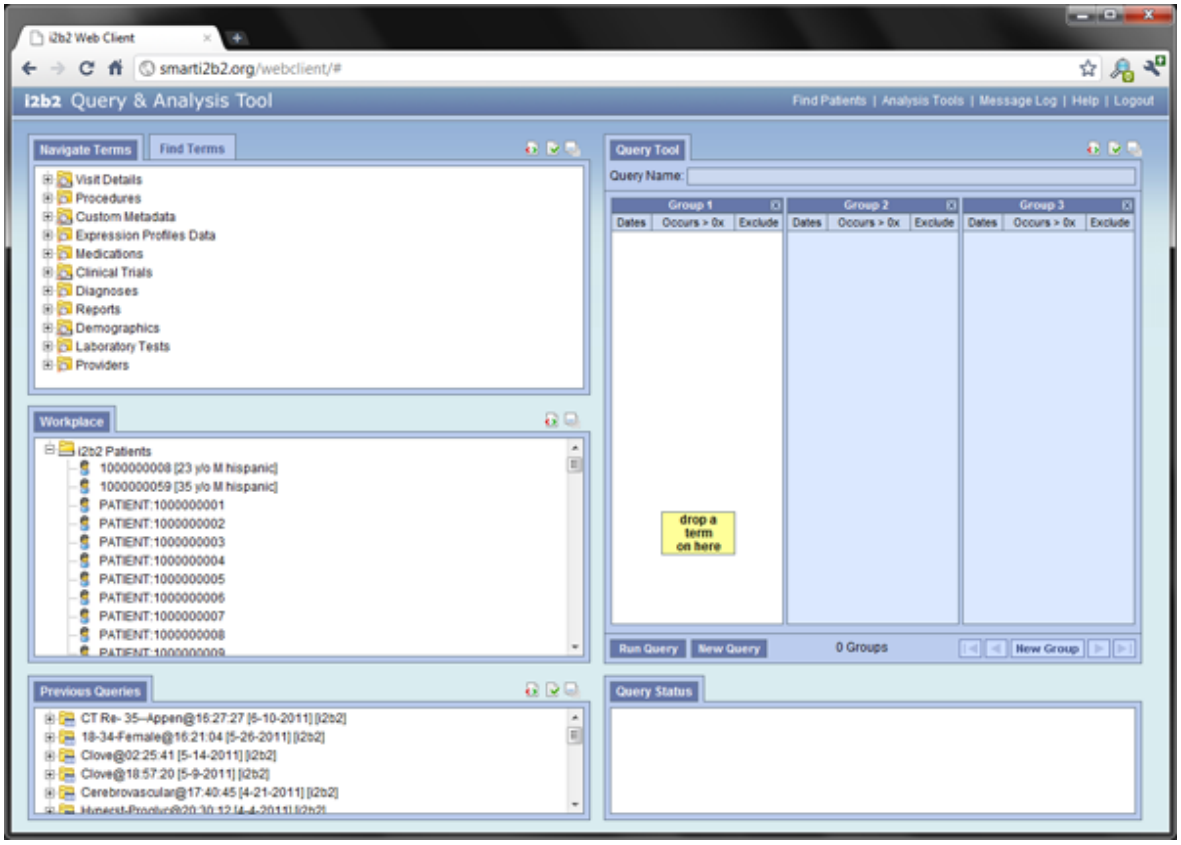

FIGURE 4.10 – Interface du workbench i2b2.

La plateforme propose l'exploitation des données soit à partir d'un logiciel client fonctionnant sous Windows soit d'une interface web, les deux outils proposant des interfaces similaires (voir FIGURE 4.10). L'interface permet de construire des requêtes en sélectionnant les concepts issus de terminologies ou d'ontologies médicales. Ces concepts recouvrent des données démographiques, diagnostiques, de prescriptions ou encore de tests médicaux. L'outil de construction des requêtes permet d'inclure plusieurs concepts et gère également l'exclusion de concepts, la fréquence d'apparition de concepts ainsi que des interrogations dans le temps. Ces fonctionnalités sont détaillées dans le tableau 4.2

La requête réalisée renvoie ensuite une cohorte de patients anonymisée correspondant aux concepts sélectionnés. Il est alors possible de visualiser ces données sous forme graphique, principalement par des représentations de type histogramme, ou d'exporter les données vers le format Excel. Le modèle i2b2 permet ainsi d'intégrer des données médicales diverses afin de pouvoir faire des interrogations multicritères dans le temps [MURPHY et al., 2006].

Bien qu'i2b2 soit un outil puissant pour gérer des données cliniques, il concentre ses efforts en particulier sur la problématique de la recherche de cohortes. Il ne permet pas la visualisation des données à l'échelle d'un seul patient et donc la visualisation des données d'un DPI [MURPHY et al., 2017]. De plus, son déploiement demande des

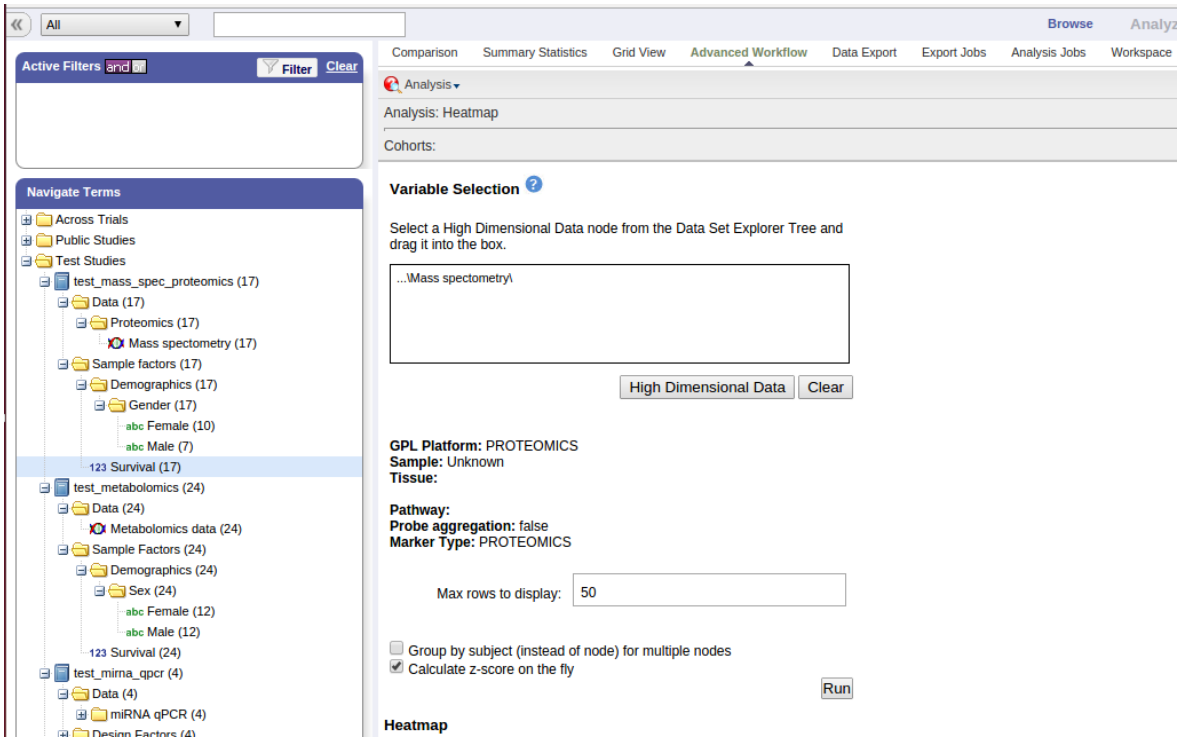

FIGURE 4.11 – Interface du client web tranSMART.

ressources importantes, en terme de déploiement, de maintenance ainsi qu'en terme de formation à son utilisation [TAKAI-IGARASHI et al., 2011].

tranSMART

tranSMART [SCHEUFELE et al., 2014; SZALMA et al., 2010] est une plateforme libre de recherche translationnelle, développée par l'entreprise pharmaceutique Johnson & Johnson et supportée par une communauté croissante de développeurs. Le modèle de données de tranSMART (version 1.2) est basé sur le modèle i2b2. La plateforme est développée en Java et repose sur plusieurs dépendances comme le logiciel R pour les analyses statistiques.

La plate-forme tranSMART permet d'intégrer des données à partir d'une variété de sources de données. Les types de données qui peuvent être intégrées dans tranSMART comprennent des données cliniques au niveau du patient ou de l'étude (démographie, diagnostic, médicaments, résultats de laboratoire), des données omiques (génotypage, expression de gènes, expression des protéines), les résultats de l'étude, ainsi que les descripteurs d'étude sous forme de métadonnées. Les données peuvent provenir de sources publiques (par exemple, TCGA, GEO), ou à partir de sources internes (par exemple, essais cliniques institutionnels, etc.). Le contenu de référence introduit dans tranSMART peut être propriétaire (par exemple, GeneGo, Ingenuity) ou disponible ouvertement (par exemple, Entrez, MeSH). En outre, les sources de données d'origine peuvent être stockées sous forme de fichiers et accessibles pour l'exportation via l'interface utilisateur

	RAVEL	**i2b2 1.7.06**
Portée	1 à n patients	n patients
Contraintes numériques		
- Par rapport à une valeur fixe	✓	✓
- Par rapport à une borne de référence	✓	✓
- Gestion des unités de mesure	✗	✓
Contraintes chronologiques		
- Par rapport à une date fixe	✓	✓
- Par rapport à un intervalle de temps	✓	✓
- Entre deux évènements	✓	✓
Contraintes textuelles	✓	✓
Recherche plein texte dans les comptes-rendus	✓	✓
Détection du nombre d'occurences d'un évènement	✗	✓
Données obtenues en sortie	Toute entité du modèle conceptuel	Set de patients
Données cliniques gérées		
- Données patient	✓	✓
- Séjours	✓	✓
- Unités médicales	✓	✓
- Analyses biologiques	✓	✓
- Actes médicaux	✓	✓
- Prescriptions	✓	✓
- Compte-rendus	✓	✓
Données omiques gérées		
- Variants :		
• SNV/SNP	✓	✓
• MNP	✓	✓
• Variants du nombre de copies	✓	✓
• Indels	✓	✓
• Inversions, substitutions	✓	✓
- Analyse d'expression :		
• Protéines	✓	✓
• Gènes	✓	✗
• micro-ARN	✓	✗
• Exons	✓	✗
- Analyse de méthylation de l'ADN	✓	✗
Terminologies prises en charge	> 50	Gene Ontology, HOM-UCARE, Human Phenotype Ontology, ICD 10, ICD 9, LOINC, MedDRA, MeSH, NDFRT, RADLEX, RxNORM, SNOMED CF+CT
Prise en charge de l'explosion	✓	✓

Tableau 4.2 – Analyse fonctionnelle des outils RAVEL et i2b2 v1.7.06.

	RAVEL	**tranSMART v16.2**
Portée	1 à n patients	n patients
Recherche possible par :		
- Étude	✓	✓
- Objectif de l'étude	✓	✓
- Type de biomarqueur	✓	✓
- Type d'analyse	✓	✓
- Technologie employée	✓	✓
- Matériel utilisé	✓	✓
- Domaine thérapeuthique	✗	✓
Contraintes textuelles	✓	✓
Données obtenues en sortie	Toute entité du modèle conceptuel	Ensemble de données
Visualisation des données cliniques	1 et n patients	✗
Visualisation des données omiques	1 et n patients	n patients
Données omiques gérées		
- Variants :		
• SNV/SNP	✓	✓
• MNP	✓	✓
• Variants du nombre de copies	✓	✓
• Indels	✓	✓
• Inversions, substitutions	✓	✓
- Analyse d'expression :		
• Protéines	✓	✓
• Gènes	✓	✓
• micro-ARN	✓	✓
• Exons	✓	✓
- Analyse de méthylation de l'ADN	✓	✓
Terminologies prises en charge	> 50	MeSH, NCBI Gene, CIM10, LOINC
Prise en charge de l'explosion	✓	✓
Fonctionnalités d'analyse	✗	aCGH Survival Analysis, Box Plot with ANOVs, Correlation Analysis, Forest Plos, Group Test for RNASes, Heatmaps, IC50 Dose Response Curvs, Line Graps, Logistic Regressios, PCs, Scatter Plot with Linear Regressios, Survival Analysis, Fisher test

TABLEAU 4.3 – Analyse fonctionnelle des outils RAVEL et tranSMART v16.2.

pour les futurs workflows de recherche.

L'interface web permet d'explorer les données phénotypiques de groupes de patients, de conduire des méta-analyses et de valider ou rechercher des hypothèses. Les outils d'analyse statistique disponibles permettent de réaliser des analyses pointues sur les données disponibles. L'explorateur de jeu de données est le principal point d'accès pour les données d'étude (voir FIGURE 4.11). Dans l'explorateur de jeu de données, l'utilisateur parcourt des données de manière hiérarchique. En utilisant le glisser-déposer, l'utilisateur peut exécuter une variété dd'analyses et statistiques. L'utilisateur peut spécifier une cohorte de patients, choisir une modalité d'analyse disponible et définir les paramètres pertinents. Une fois les sélections effectuées, les résultats sont retournés à l'utilisateur et présentés de manière graphique. L'utilisateur peut continuer à explorer davantage les données en modifiant les détails de la cohorte ou les paramètres de l'analyse ou en exécutant d'autres analyses. Ainsi, l'utilisateur peut générer une hypothèse dans tranSMART en utilisant un ensemble de données et ensuite tester cette hypothèse sur un ensemble de données différent. Les résultats générés peuvent être sauvegardés ou exportés. Aucune connaissance des langages de script statistique n'est nécessaire pour analyser les données. Ces fonctionnalités sont détaillées dans le tableau 4.3.

La plateforme tranSMART est ainsi dédiée en particulier à l'exploration analytique d'ensemble de données omiques plutôt qu'à la visualisation et la récupération d'information dans les données des DPI. Elle ne permet aucun accès aux données d'un seul patient, ni en visualisation ni en recherche d'information au contraire de la solution RAVEL. En revanche, elle propose des fonctionnalités d'analyse statistiques des données puissantes et aisée d'accès qui ne sont pas disponible dans notre application.

4.3.2 Cas clinique RAVEL en rhumatologie : polyarthrite rhumatoïde

Objectifs du cas d'usage

La polyarthrite rhumatoïde (PR) est une maladie inflammatoire auto-immune à prédominance synoviale (articulation). Elle évolue par poussées et peut entraîner des lésions articulaires graves. C'est la plus fréquente des rhumatismes inflammatoires chroniques, avec une prévalence entre 0.4 et 0.8% en France.

La présentation initiale de la maladie est hétérogène, le diagnostic peut parfois s'avérer difficile avec un laps de temps avant la découverte de la maladie. À l'inverse, un diagnostic de PR peut être remis en cause après plusieurs échecs thérapeutiques par exemple : Le clinicien aimerait alors avoir un résumé des éléments objectifs ayant conduit à ce diagnostic pour éventuellement l'infirmer.

Une fois le diagnostic établi, le patient est régulièrement suivi, en consultation pour les cas les moins graves, en hôpital de jour pour les cas plus sévères où le traitement doit

être administré en injection. La majorité du suivi se fait via un formulaire spécifique.

Il arrive parfois qu'un patient doive être hospitalisé en service traditionnel (forme grave, échec thérapeutique, effet indésirable grave) où les formulaires ne sont pas utilisés.

Il s'agit de retrouver dans le dossier les éléments permettant de vérifier la réalité du diagnostic de PR (remise en cause du diagnostic après plusieurs échecs thérapeutiques par exemple) :

- Les traitements : leurs effets indésirables, et en particulier le nombre de polynucléaires neutrophiles (PNN). Pouvoir différentier :
  - nne neutropénie brutale et inhabituelle (Neutropénie : Taux bas, inférieur à 1 500 par mm3, de PNN- (granulocytes, type de globules blancs) ;
  - une neutropénie cyclique.
- Évolution de variables biologiques marqueurs de l'inflammation : CRP, VS, FR, Anti-CCP ;
- Évolution du DAS-28 (score d'activité de la maladie) ;
- Évolution des atteintes articulaires.

Éléments de recherche

Les éléments à rechercher dans ce cas d'usage sont les suivants :

1. Quelles sont les articulations gonflées pour chaque visite ?
2. Quelles sont les articulations douloureuses pour chaque visite ?
3. Quelles sont les articulations gonflées et douleurs pour chaque visite ?
4. Combien de grosses articulations sont atteintes à chaque visite ?
5. Combien de petites articulations ?
6. Sérologies RF (valeur + normal ou pas ?)
7. ACPA (valeur + normal ou pas ?)
8. VS (valeur + normal ou pas ?)
9. CRP (valeur + normal ou pas ?)
10. Durée des symptômes
11. Traitements (switch inefficacité ou EIM)
12. DAS-28
13. PNN

Tableau 4.4 – Réponse au cas d'usage PR RAVEL.

Élement	Nombre de réponses	Temps d'exécution
1	112	3.17s
2	6659	1.14s
3	76	0,37s
4	34 entrées	0.89s
5	17 entrées	34.55s
6	17 entrées	39.52s
8	8 entrées	4.12s
9	11 entrées	1.52s
13	5 entrées	1.47s

Résultats

RAVEL a permis d'obtenir les éléments concernant les articulations (1-5), la sérologie RF (6), la VS (8), la CRP (9) ainsi que le PNN (13). Les éléments concernant les articulations ont nécessité d'utiliser la SNOMED CT. En effet, la caractéristique de l'articulation, petite ou grosse, n'étant pas codée dans les questionnaires mais seulement l'articulation elle-même, il a été nécessaire de parcourir les concepts ascendants pour déterminer la caractéristique de l'articulation touchée. Ceci explique un temps de réponse plus important pour ces éléments. Les valeurs de tests biologiques comme le taux de CRP ont pû être recherchées en un temps satisfaisant. Plusieurs éléments n'ont pas pu être retrouvés. Les valeurs ACPA n'ont pû être intégrées dans la terminologie locale et n'ont donc pas pû être récupérée. Les éléments de questionnaires comme la durée des symptômes ou la valeur DAS-28 étaient présents dans des champs de texte libre de questionnaires et n'ont pû être directement interrogés.

4.4 Synthèse

Dans ce chapitre, nous avons vu en premier lieu l'intégration des données de références issues des bases de connaissances NCBI Gene, Uniprot SwissprotKB et OMIM. Ces données mises à jour régulièrement ont été intégrées au système d'information du D2IM du CHU de Rouen et sont ainsi accessibles d'une part au public via le portail terminologique HeTOP et d'autre part pour notre application de RI dans les données cliniques et omiques. L'intégration de données expérimentales a nécessité l'intégration de jeux de données de natures, de formats, et de sources diverses. Au total, neuf études ont été intégrées couvrant les types de données omiques utilisés en recherche clinique : des données d'expression de gènes, protéines, exons et microARN, des variants génétiques, et enfin des analyses de méthylation de l'ADN. Par la suite, en m'appuyant sur

les travaux réalisés pour les données cliniques dans le cadre du projet RAVEL et sur le moteur réalisés par l'équipe du D2IM du CHU de Rouen, j'ai adapté le moteur de recherche existant aux données omiques pour aboutir à un outil permettant de requêter à la fois des données cliniques et omiques, à l'échelle d'un ou plusieurs patients. Une interface graphique a également été conçue pour la visualisation des données omiques. Enfin, l'évaluation de ce travail face aux plateformes i2b2 et TranSMART a été réalisée. L'évaluation de la partie clinique de notre solution à travers un cas d'usage mis au point dans le cadre du projet RAVEL a également été décrite. Ce travail a fait l'objet de plusieurs communications [Cabot et al., 2016c, 2015]

Plusieurs points peuvent être discutés. Tout d'abord, l'intégration de multiples sources de données omiques expérimentales a permis de valider la pertinence du modèle de données omiques développé. Cependant, il n'a pas été possible dans le cadre de cette thèse de disposer d'un jeu de données cliniques et omiques d'un nombre suffisant de patients. Notre prototype se base donc aujourd'hui sur des données cliniques et omiques décorrélées : les données cliniques provenant de DPI du CHU de Rouen et les données omiques provenant de sources diverses récoltées sur des patients différents. Bien qu'une certaine cohérence ait pu être conservée en s'appuyant sur les codes diagnostic CIM10 pour lier les données entre elles, cette décorrélation est un frein important à l'évaluation du système. L'évaluation du système en situation réelle, par l'équipe d'un laboratoire sur ces propres données par exemple, n'a ainsi pas pu être effectuée et des points importants comme l'ergonomie, la pertinence des vues proposées ou l'apprentissage de l'écriture des requêtes ne sont pas évalués. Le prototype réalisé permet néanmoins de démontrer la validation technique du système et son évaluation complète reste une perspective forte de ce travail.

Enfin, face à l'existant, notre solution conserve l'avantage de proposer au sein d'un système unique de permettre l'interrogation conjointe de données cliniques et données omiques à l'échelle d'un ou plusieurs patients. Ses fonctionnalités sont ainsi étendues : il peut être utilisé dans le cadre de la pratique clinique comme dans le cadre de la recherche clinique pour la création de cohortes et plus largement en santé publique et épidémiologie. L'utilisation d'un système unique permet d'une part de concentrer les efforts de développement et de maintenance, mais aussi du point de vue de l'utilisateur d'avoir accès à ces fonctionnalités multiples sans devoir maîtriser plusieurs systèmes. Enfin, ce choix favorise également la non-redondance des informations et la continuité des connaissances entre recherche et pratique clinique indispensable à la personnalisation des soins.

Chapitre 5

Indexation multi-terminologique de documents biomédicaux

Sommaire

Dans ce chapitre, il s'agit de traiter l'indexation multi-terminologique de documents biomédicaux. L'équipe du D2IM du CHU de Rouen développe l'ECMT. Il exploite les terminologies intégrées au portail HeTOP développé dans le cadre du projet PlaIR. L'utilisation de multiples terminologies pour l'indexation pose des problématiques spécifiques. Ce chapitre présente une méthode de priorisation des indexations visant à traiter la problématique du bruit induit par la multi-terminologie. Il présente enfin deux évaluations. En effet, l'outil ECMT a été évalué à deux reprises dans le cadre de la campagne de test CLEF eHealth. Une évaluation de la couverture terminologique au sein d'un corpus issu de la base de données en français LiSSa est également présentée.

5.1 Le serveur multi-terminologique HeTOP

Il y a un intérêt croissant aujourd'hui, non seulement pour développer et maintenir des terminologies de soins de santé, mais aussi pour les rendre interopérables dans les systèmes d'information fournissant des services aux applications [Soualmia et Darmoni, 2005]. Un « serveur de terminologie » est un outil qui gère et donne accès à plusieurs terminologies [Burgun et al., 1997]. De nombreux serveurs terminologiques ont déjà été développés, principalement en anglais [Brown et al., 2007; Burgun et al., 1997; Gambarte et al., 2007; Komatsoulis et al., 2008; Navas et al., 2007]. L'objectif principal d'un serveur multi-terminologique et multilingue est de créer une base de données terminologique permettant de rechercher des concepts et des termes parmi des ressources terminologiques et les parcourir dynamiquement. De telles données peuvent être utilisées pour (i) indexer les ressources manuellement ou automatiquement, (ii) permettre la recherche d'information à partir de plusieurs terminologies, (iii) évaluer l'intégrité des données terminologiques et (iv) fournir des ressources éducatives.

5.1.1 Un serveur multiterminologique et interlingue

Les données biomédicales se développent constamment, en particulier avec les nouvelles technologies et les médias Internet. Par conséquent, il devient obligatoire d'indexer et annoter ces données avec des vocabulaires contrôlés et structurés afin de stocker et rechercher ces informations avec des méthodes intelligentes. Un aspect clé de l'interopérabilité sémantique pour les données dans les sciences de la vie est l'utilisation de terminologies ou d'ontologies comme dénominateur commun pour structurer les données et les rendre interopérables [Bodenreider et Stevens, 2006; Rubin et al., 2008]. Les terminologies et ontologies sont des classifications qui décrivent la connaissance d'un domaine spécifique avec des concepts et des relations entre eux. Les ontologies sont plus complexes que les terminologies, car elles peuvent définir des règles et des fonctions pour inférer et structurer les connaissances. Étant donné que les terminologies et ontologies sont couramment utilisées dans les systèmes d'information et

spécifiquement en santé, il est difficile de les gérer et de les consulter au sein d'une même application. En effet, de nombreuses terminologies et ontologies ont été créées au cours de la dernière décennie à des fins différentes : l'indexation et l'annotation de documents, l'organisation des connaissances, l'inférence des faits, etc. Certaines terminologies et ontologies (par exemple MeSH, CIM10, CCAM) sont couramment utilisées quotidiennement dans les hôpitaux ou dans les laboratoires de recherche et sont très utiles pour la recherche d'information. Les terminologies et ontologies ne sont pas toujours bien structurées ou définies en raison d'un manque de standardisation ou de mise en forme. En outre, la sémantique et l'interopérabilité syntaxique entre terminologies et ontologies sont un grand défi pour permettre l'interconnexion entre les systèmes et les connaissances. Plusieurs outils ont été créés pour stocker, rechercher et utiliser plusieurs terminologies et ontologies en même temps : parmi eux, l'UMLS (Unified Medical Language System) décrit dans la section 2.3.1, le service de recherche d'ontologies EBI [Côté et al., 2010], le NCBO BioPortal [Rubin et al., 2006] et le portail HeTOP [Grosjean et al., 2011].

HeTOP (Health Terminology Ontology Portal) est un portail de ressources terminologiques et ontologiques développé au sein de l'équipe CISMeF (LITIS EA 4108 - TIBS) dans le cadre de la thèse de Julien Grosjean [Grosjean, 2014]. Il héberge 69 terminologies et ontologies dans plusieurs langues. La plupart des terminologies et ontologies sont des références nationales ou internationales telles que le MeSH, la CIM10 ou la CCAM. Ces terminologies et ontologies sont régulièrement mises à jour et sont accessibles via un site Web [1] et un service Web. HeTOP a été conçu comme un portail multi-terminologique de référence et un portail interlingue pour aider les libraires, les traducteurs, les étudiants et les professionnels de la santé à récupérer des ressources et des connaissances dans une grande variété de domaines médicaux complexes. Il ne faut pas confondre interlinguisme (en anglais cross-lingual) qui désigne le fait de passer d'une langue à une autre en conservant le sens (au maximum) avec le multilinguisme qui correspond au fait de gérer plusieurs langues dans un système donné. Ainsi, HeTOP est à la fois interlingue, car il permet de passer d'une langue à une autre via les concepts, mais il est également multilingue, car il propose de rechercher des termes dans plusieurs langues à la fois. L'interface graphique offerte par le site web est traduite complètement ou partiellement en 12 langues. À l'instar de la NCBO, CISMeF développe des outils et des services supplémentaires pour utiliser et exploiter ces ressources dans HeTOP : (i) l'ECMT [2] effectue l'indexation des documents, identifiant les concepts de terminologies et ontologies présents dans un texte [Cabot et al., 2016b; Pereira et al., 2008; Soualmia et al., 2015], (ii) InfoRoute [3], qui est un service d'info-bouton permettant d'accéder à de nombreux portails à partir d'une requête simple [Darmoni

1. www.hetop.org
2. http://ecmt.chu-rouen.fr/
3. http://inforoute.chu-rouen.fr/

et al., 2008], en effectuant une extension sémantique basée sur les terminologies et ontologies, (iii) MT@HeTOP [4] est un service basé à la traduction et l'alignement de termes, (iv) Doc'CISMeF [5] [Darmoni et al., 2001] qui est un moteur de recherche de ressources Web de qualité concernant la santé manuellement ou automatiquement mis à jour par les documentalistes de CISMeF avec le MeSH et d'autres terminologies et ontologies de référence.

5.1.2 Terminologies et ontologies disponibles

Le portail HeTOP gère 69 terminologies en français et en anglais, totalement ou entièrement traduites en français, alignées par des relations sémantiques. Dans sa dernière version, le système de gestion de base de données relationnelle est remplacé par le système d'analyse Infinispan pour permettre un traitement rapide des entrées. Les principaux objectifs sont l'optimisation des temps de réponse et la dissociation du moteur de recherche d'un système de gestion de bases de données propriétaires. La solution NoSQL Infinispan permet la distribution et la récupération des données à partir de plusieurs serveurs. Certaines de ces ressources sont issues du Métathésaurus UMLS. À ce jour, les principales ressources disponibles sont :

- CIM10 dans les versions Organisation mondiale de la Santé (OMS) et ATIH, pour les diagnostics ;
- CCAM, pour les actes médicaux ;
- Medline Plus ;
- MeSH, incluant les concepts supplémentaires ;
- SNOMED CT et la version internationale ;
- Anatomical Therapeutic Chemical classification (ATC), pour les médicaments ;
- MEDical Dictionary for Regulatory Activities terminologies (MedDRA) ;
- MedDRA
- Foundational Model of Anatomy (FMA)
- Human Phenotype Ontology (HPO)
- LOINC
- National Cancer Institute thesaurus (NCIt)

Le Tableau A.1 contient leur volumétrie. Chaque concept de ces ressources, lorsqu'il est disponible dans l'UMLS, possède un identificateur unique de concept. C'est le cas par exemple pour la CIM10 et non pour la CCAM.

4. http://cispro.chu-rouen.fr/MT_EHTOP/
5. http://doccismef.chu-rouen.fr

5.2 L'Extracteur de Concepts Multi-Terminologique (ECMT)

Dans le cadre des travaux liés à l'évaluation des systèmes d'information sur la santé et la recherche et l'indexation de l'information dans le DPI [Cabot et al., 2016a; Lelong et al., 2016], un outil nommé ECMT est développé par le D2IM du CHU de Rouen. Il a été utilisé dans plusieurs projets subventionnés par l'Agence Nationale de la Recherche [Dupuch et al., 2013; Thiessard et al., 2012]. Pour évaluer les performances de l'ECMT, notre équipe a participé pour la première fois à la compétition CLEF eHealth en 2015 [Goeuriot et al., 2015], puis en 2016 [Kelly et al., 2016]. La principale motivation de cette participation est d'améliorer les fonctionnalités de l'outil. La tâche de reconnaissance d'entités cliniques est retenue [Névéol et al., 2016, 2015]. Elle vise à identifier automatiquement les entités cliniques pertinentes dans des textes médicaux en français. L'ECMT utilise le traitement du langage naturel, la recherche de motifs et exploite plusieurs terminologies et ontologies pour réaliser cette reconnaissance.

5.2.1 Détection des concepts

L'ECMT est développé pour extraire des textes en entrée une liste des concepts de santé candidats des 69 terminologies et ontologies incluses dans HeTOP. L'extraction est effectuée au niveau de la phrase du texte. Un service Web SOAP et REST permet de fournir une réponse en XML pour chaque concept qui contient : l'indice du premier et du dernier mot qui a conduit à identifier le concept médical dans la liste finale, l'identifiant et son type sémantique si le concept de santé est inclus dans le Métathesaurus UMLS et la spécialité médicale du concept. Ces derniers sont basés sur des liens sémantiques manuels entre les spécialités médicales générales (par exemple, la dermatologie, l'oncologie, etc.) et les terminologies et ontologies incluses dans HeTOP. L'ECMT s'appuie sur l'algorithme du sac de mots et également la reconnaissance de motifs pour analyser des résumés de décharge, des rapports de procédures ou des résultats de laboratoire qui contiennent des données symboliques (présence ou absence), des données numériques et des unités de mesure. La méthode du sac de mots a été développée principalement pour la RI et elle a été adaptée pour l'indexation, c'est-à-dire que seul le plus grand ensemble de mots qui correspond à un concept est extrait, même si les sous-ensembles proposent d'autres concepts. Cette méthode est considérée comme étant plus précise et évite le bruit. Le texte en entrée est normalisé et chaque phrase est traitée séparément pour extraire les concepts. L'ECMT dispose également d'une interface conviviale (figure 1) accessible après authentification [6]. Plusieurs options sont

6. http://ecmt.chu-rouen.fr/

disponibles pour paramétrer l'indexation du texte :

- `c` : catégorisation. Si `c=Vrai`, les spécialités médicales ainsi que le type sémantique du concept sont inclus dans la réponse (valeur par défaut : `Vrai`).
- `r` : restriction. Si `r=Vrai`, la recherche s'arrête quand un concept correspondant à un maximum de mots est trouvé (valeur par défaut : `Vrai`). Par exemple, pour « cardiopathie hypertensive », si `r=Vrai`, seul le concept *hypertension artérielle* est retourné. Si `r=Faux`, la méthode retourne les concepts *hypertension artérielle* et *maladie cardiaque*.
- `sn` : réseau sémantique. Si `sn=Vrai` les concepts alignés aux concepts trouvés dans le texte sont inclus dans la réponse (valeur par défaut : `Faux`).
- `e` : exclusions. Il s'agit d'une chaîne de caractères contenant les identifiants des concepts à exclure de la réponse (une spécialité médicale, un type sémantique, un ancêtre, etc.). Par exemple, `e=CIS_MT_8,UML_ST_T060, MSH_D_C` retourne uniquement les concepts qui ne sont pas des chirurgies (`CIS_MT_8`), ni des procédures de diagnostic (`UML_ST_T060`) ou des maladies MeSH (`MSH_D_C`). Par défaut, toutes les catégories sont retournées. Si un ancêtre est utilisé pour l'exclusion, tous ces descendants sont exclus de la réponse.
- `f` : filtres. De la même façon que pour les exclusions, il s'agit d'une chaîne de caractères contenant les identifiants des concepts à conserver uniquement dans la réponse.
- `a` : ancêtres. Si `a=Vrai`, l'ECMT retourne les ancêtres de chaque concept trouvé dans sa réponse (valeur par défaut : `Faux`).
- `d` : descendants. Si `d=Vrai`, l'ECMT retourne également les descendants de chaque concept trouvé dans sa réponse (valeur par défaut : `Faux`).
- `at` : synonymes. Si `at=Vrai`, les synonymes de chaque concept trouvé sont inclus dans la réponse (valeur par défaut : `Vrai`).

La réponse du service Web est un fichier XML qui sérialise la sortie de l'annotation du texte. Les champs suivants le composent :

- `<cis-sentences>` : le texte en entrée ;
- `<timemillis>` : temps d'exécution en millisecondes ;
- `<cis-sentence>` : une phrase ;
- `<idsentence>` : l'identifiant de la phrase ;
- `<position>` : indice de début de la phrase dans le texte ;
- `<start>` : indice de début de l'indexation ;
- `<end>` : indice de fin de l'indexation ;
- `<idterm>` : identifiant d'origine du concept ;

— `<offset>` : indice des termes composant le concept ;
— `<ter>` : identifiant de la terminologie du concept ;
— `<umlscui>` : CUI ;
— `<matchterms>` : termes ayant permis d'identifier le concept ;
— `<cis:term>` : libellé préféré du concept ;
— `<cis:label>` : libellé ;
— `<lang>` : langage du libellé ;
— `<cis:altterms>` : liste des synonymes du concept ;
— `<cis:altterm>` : synonyme du concept ;
— `<cis:categorization>` : liste de spécialités médicales ou types sémantiques du concept ;
— `<cis:category>` : spécialité médicale ou type sémantique du concept ;
— `<cis:descendants>` : liste des descendants du concept ;
— `<cis:descendant>` : descendant du concept ;
— `<cis:ancestors>` : liste des ancêtres du concept ;
— `<cis:ancestor>` : ancêtre du concept ;
— `<cis:relateds>` : liste des concepts reliés sémantiquement au concept trouvé ;
— `<cis:related>` : concept relié sémantiquement au concept trouvé ;
— `<relationLabel>` : libellé de la relation.

5.2.2 Exploitation des réseaux sémantiques

Une nouvelle option nommée `prioritisation` a été ajoutée depuis 2015. Elle traite le problème spécifique lié au bruit généré par l'indexation multi-terminologique. Si cette option est activée, l'ECMT renvoie uniquement le concept le plus fiable, selon son type sémantique (valeur par défaut : `Faux`).

Lorsque n termes identiques provenant de plusieurs terminologies sont récupérés, les types sémantiques liés à ces termes sont déterminés et le plus important est déterminé à l'aide d'opérations ensemblistes. Ensuite, le terme le plus pertinent est conservé en fonction d'une classification des ressources HeTOP conçues manuellement pour chaque type sémantique disponible dans l'UMLS. Par exemple, l'indexation du terme « asthme » avec l'ECMT génère sept concepts récupérés dans sept ressources différentes : Systematized Nomenclature of Medicine (SNOMED Int.), NCIt, MeSH, Medline Plus, HPO, CIM10 et International Classification for Nursing Practice (ICNP). Avec l'option de priorisation activée, un seul concept est récupéré selon le type sémantique correspondant au concept *asthme* dans l'UMLS (maladie `T47` dans ce cas) qui est un concept

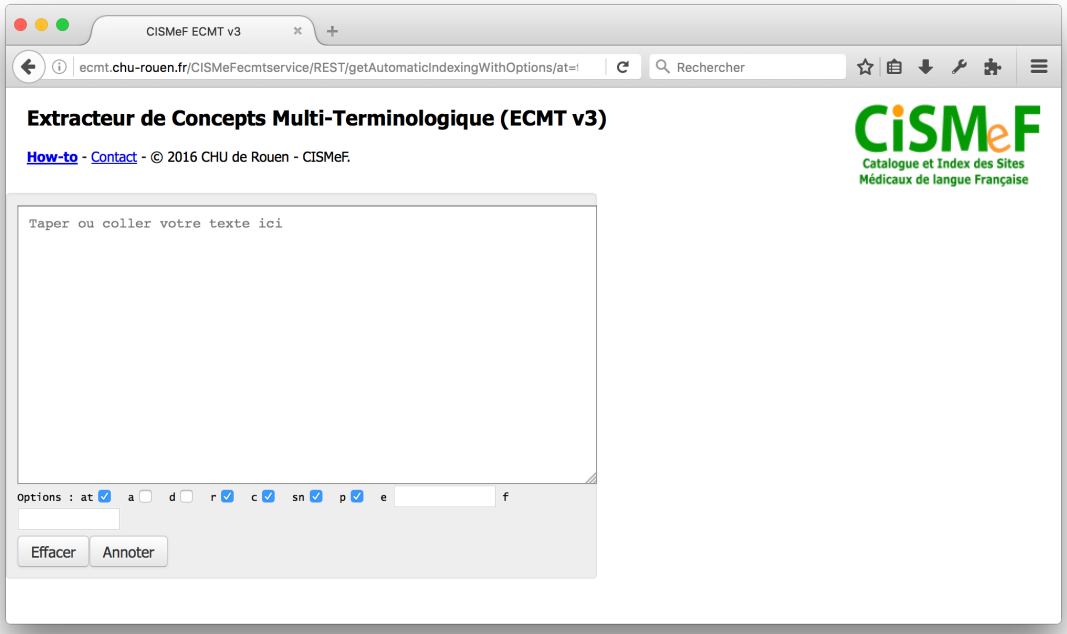

FIGURE 5.1 – L'interface utilisateur de l'ECMT et ses options.

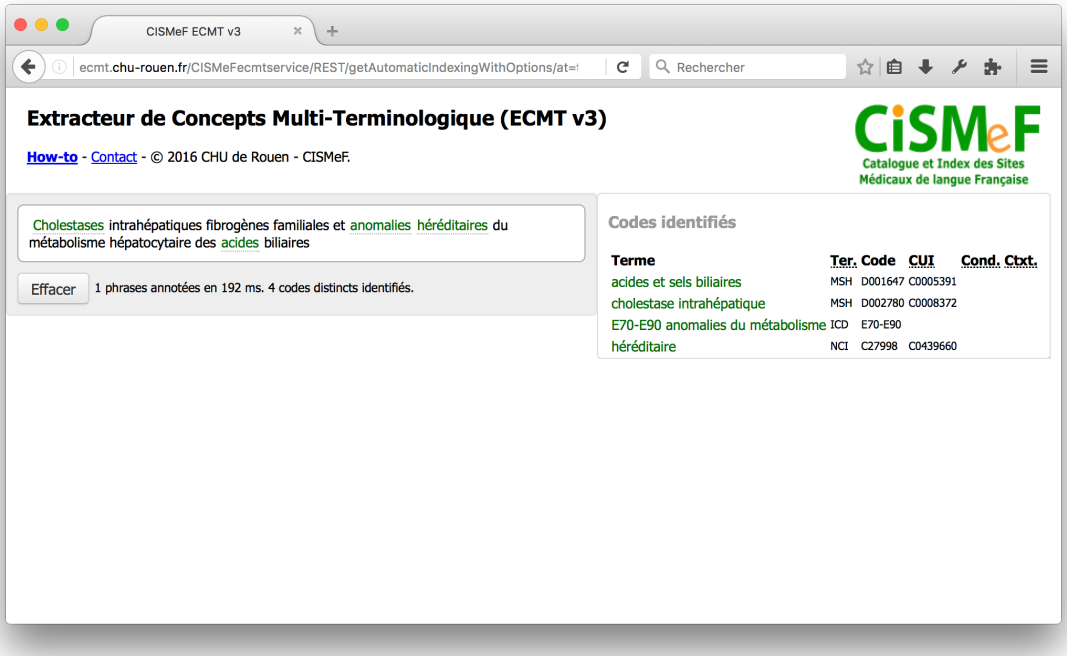

FIGURE 5.2 – Exemple du traitement de la phrase « Cholestases intrahépatiques fibrogènes familiales et anomalies héréditaires du métabolisme hépatocytaire des acides biliaires » avec l'ECMT et l'option de priorisation activée.

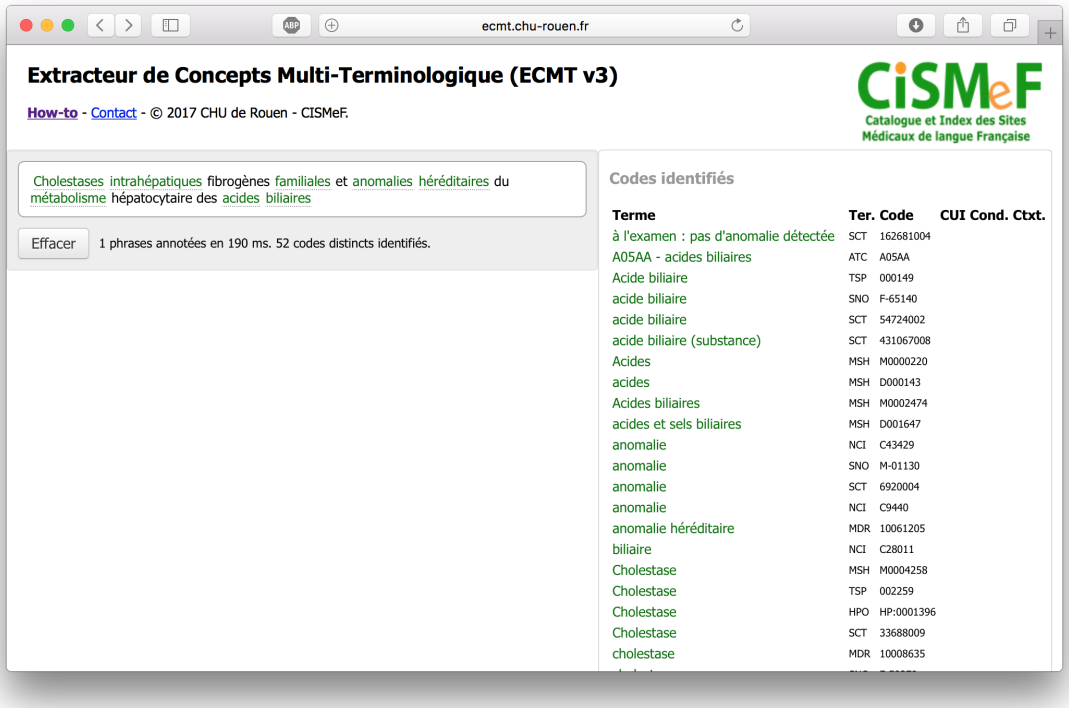

FIGURE 5.3 – Exemple du traitement de la phrase « Cholestases intrahépatiques fibrogènes familiales et anomalies héréditaires du métabolisme hépatocytaire des acides biliaires » avec l'ECMT et l'option de priorisation désactivée.

MeSH. Si aucun concept MeSH ne pouvait être récupéré pour un concept de maladie `T47`, un concept NCIT devrait être priorisé et récupéré, et ainsi de suite. Cette option permet donc‘′ de conserver le concept de la terminologie la plus pertinente pour le type sémantique considéré.

Les résultats de cette option sont illustrés dans les FIGURE 5.2 (priorisation activée) et 5.3 (priorisation désactivée). Elles donnent un exemple de traitement de la phrase « Cholestases intrahépatiques fibrogènes familiales et anomalies héréditaires du métabolisme hépatocytaire des acides biliaires » avec toutes les options par défaut de l'ECMT et l'option de priorisation activée. L'ECMT extrait les termes MeSH *acides et sels biliaires* (CUI C0005391), *cholestase intrahépatique* (CUI C0008372), le terme CIM10 *E70-E90 anomalies du métabolisme* et le terme NCIt *héréditaire* (CUI C0439660). L'utilisateur peut également visualiser les termes et catégories synonymes.

5.3 Évaluation de l'indexation au sein des corpus MEDLINE et EMEA

Les performances de l'ECMT ont été évaluées à l'occasion des éditions 2015 et 2016 de la compétition CLEF eHealth [CABOT et al., 2016b; SOUALMIA et al., 2015] dans les tâches correspondantes à l'extraction d'information dans des textes médicaux.

L'édition 2015 a permis de dresser un premier état des lieux des performances de cet outil puisqu'il s'agissait de sa première évaluation dans le cadre d'une compétition internationale. L'édition 2016 a permis d'évaluer les progrès réalisés dans la reconnaissance d'entités et plus particulièrement, dans la gestion des spécificités de l'indexation multi-terminologique.

5.3.1 Description des tâches

Reconnaissance d'entités nommées La tâche de la reconnaissance d'entité nommée consiste à analyser des documents texte afin de marquer les dix types d'entités d'intérêt clinique définis dans la compétition (voir la section 2.1). Deux sous-tâches sont possibles : (i) reconnaissance des entités simples (extraction du concept et des indices du texte correspondant) et (ii) reconnaissance des entités normalisées (extraction du concept et de son CUI et des indices du texte correspondant).

Normalisation des entités La tâche de normalisation des entités consiste à identifier pour chaque entité d'intérêt extraite le CUI UMLS correspondant.

5.3.2 Sources de données

Le corpus QUAERO

Le corpus QUAERO a été développé en tant que ressource pour la reconnaissance et la normalisation des entités nommées en 2013 [Neveol et al., 2014a] dans le cadre du défi 2013 de CLEF-ER, dans le but de créer un ensemble standard d'entités normalisées pour l'analyse des textes biomédicaux en français. Une sélection des titres issus de la base MEDLINE et de fiches descriptives de médicaments issues de l'Agence Européenne du Médicament (EMEA) utilisés dans le défi CLEF-ER 2013 ont été sélectionnés pour l'établissement d'un gold standard. Des annotations sont fournies dans le format BRAT [7] et le processus d'annotation a été guidé par les concepts de l'UMLS. Dix types d'entités cliniques qui sont des groupes sémantiques UMLS ont été annotées : Anatomie, chimie et drogues, dispositifs, troubles, zones géographiques, êtres vivants, objets, phénomènes, physiologie, procédures. Les annotations ont été faites de manière globale, de sorte que les entités imbriquées ont été considérées, et ainsi une entité peut correspondre à plusieurs CUI UMLS. En particulier, (i) si une mention peut se référer à plus d'un groupe sémantique, tous les groupes sémantiques pertinents devraient être annotés. Par exemple, la mention « récidive » dans la phrase « prévention des récidives » devrait être annotée avec le groupe *Maladie* (CUI C2825055) et le groupe *Phénomène* (CUI C0034897), (ii) si une mention peut se référer à plus d'un concept UMLS dans le même groupe sémantique, tous les concepts pertinents devraient être annotés. Par exemple, la mention « obsessionnels » dans la phrase « patients obsessionnels » devrait être annotée avec les CUI C0564408 et C0338831 (groupe *Trouble*) et (iii) les entités imbriquées devraient être annotées. Par exemple, dans l'expression « infarctus du myocarde », la mention « myocarde » devrait être annotée avec le groupe *Anatomie* (CUI C0027061) et la mention « Infarctus du myocarde » devrait être annotée avec le groupe *Trouble* (CUI C0027051).

L'outil d'évaluation BRATEval

L'outil d'évaluation BRATEval effectue une comparaison par paire des jeux d'annotations obtenus sur un même ensemble de documents. Les jeux annotés doivent être formatés dans le format d'annotation de BRAT [8]. La version actuelle de l'outil a été testée sur les annotations réalisées avec Brat v1.3. L'outil n'a besoin que du fichier jar `brateval.jar` pour fonctionner, qui est inclus dans le fichier de distribution, et aucune autre bibliothèque n'est requise. La performance du système a été évaluée par les mesures habituelles de l'extraction d'information : précision, rappel et mesure F1 pour la reconnaissance d'entités et la normalisation des entités telles qu'elles ont été

7. http://brat.nlplab.org/standoff.html
8. http://brat.nlplab.org/standoff.html

définies dans la section 2.7.

Les mesures de la performance ont été calculées au niveau du document et sur l'ensemble du corpus. La performance du système est obtenue en comparant les sorties du système concerné aux annotations standard de référence sur le jeu de données de test à l'aide de BRATEval. Pour la reconnaissance d'entités, une correspondance exacte (exact match) a été comptée lorsque le type d'entité et les indices fournis correspondaient à la référence. Pour la reconnaissance d'entités normalisées, une correspondance exacte a été comptée lorsque le type d'entité, les indices fournis et les CUI du système correspondaient à la référence. Pour la tâche de normalisation, les correspondances ont été comptées pour chaque CUI fourni avec une entité. En conséquence, si le système ou la référence fournissait plusieurs CUI pour une entité, un crédit partiel a été attribué au système évalué pour chaque CUI commun avec la référence.

5.3.3 Résultats CLEF e-Health 2015

Pour chaque jeu de données MEDLINE et EMEA, le service Web de l'ECMT est exécuté avec les options `restriction`, `catégorisation`, `réseau sémantique` activées pour les tâches de reconnaissance d'entités et de reconnaissance d'entités normalisées. Pour satisfaire les conditions d'évaluation de la tâche, la sortie de l'ECMT est convertie du format XML en format BRAT. La Figure 5.4 est un exemple de fichier d'annotation obtenu lors de l'indexation de la phrase « L' hyperplasie médullosurrénalienne : une étiologie rare de l' hypertension artérielle – rapport d' un cas ».

```
T1 DISO 3 35 hyperplasie médullosurrénalienne
#1 AnnotatorNotes T1 C0020507
T2 DISO 63 86 hypertension artérielle
#2 AnnotatorNotes T2 C0020538
T3 ANAT 76 86 artérielle
#3 AnnotatorNotes T3 C0003842
```

Figure 5.4 – Fichier d'annotation au format BRAT contenant des entités extraites par l'ECMT.

Les résultats obtenus par l'ECMT pour chacun des jeu de données EMEA et MEDLINE pour les tâches de reconnaissance d'entités et de reconnaissances d'entités normalisées sont présentés dans les tableaux 5.1, 5.2, 5.3 et 5.4.

Les résultats obtenus dans cette première évaluation ne sont pas satisfaisants, en particulier pour le corpus EMEA où nous obtenons pour la reconnaissance d'entités une précision de 0,0040 et un rappel de 0,0022 en exact match et une précision de 0,4345 et un rappel de 0,2986 en inexact match. Pour la reconnaissance d'entités normalisées, nous obtenons en exact match une précision de 0,0044 et un rappel de 0,0024 et en inexact match une précision de 0,2305 et un rappel de 0,1440.

TABLEAU 5.1 – Résultats obtenus lors de la compétition CLEF eHealth 2015 avec l'ECMT - QUAERO Phase 1 (EMEA) - Reconnaissance d'entités.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT	0,0040	0,0022	0,0028	0,4345	0,2986	0,3539
Moyenne	0,30912	0,3284	0,3108	0,4815	0,5198	0,4880
Médiane	0,2117	0,1835	0,2242	0,5767	0,5500	0,5538

TABLEAU 5.2 – Résultats obtenus lors de la compétition CLEF eHealth 2015 avec l'ECMT - QUAERO Phase 1 (EMEA) - Reconnaissance d'entités normalisées.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT	0,0044	0,0024	0,0031	0,2305	0,1440	0,1773
Moyenne	0,2854	0,2738	0,2792	0,4245	0,43218	0,4230
Médiane	0,0044	0,0071	0,0047	0,4234	0,5817	0,4901

Dans le corpus MEDLINE, pour la reconnaissance d'entités nous obtenons une précision de 0,2284 et un rappel de 0,1335 en exact match, une précision de 0,7091 et un rappel de 0,6366 en inexact match. Pour la reconnaissance d'entités normalisées, nous obtenons une précision de 0,2953 et un rappel de 0,1861 en exact match, une précision de 0,5003 et un rappel de 0,3638 en inexact match. Les mauvais résultats obtenus pour le corpus MEDLINE peuvent être expliqués par les duplicats existants dans les terminologies et ontologies utilisées qui diminuent la précision et par les concepts issus de ressources non incluses dans l'UMLS et pour lesquelles aucun CUI ni groupe sémantique n'est disponible, induisant une augmentation du bruit. En outre, les résultats en correspondance exacte, par rapport aux résultats en correspondance inexacte, pourraient s'expliquer par de légères différences de termes utilisés. Le gold standard utilise des libellés UMLS tandis que l'ECMT utilise les libellés préférés de la terminologie d'origine. Cela entraîne des différences mineures entre le gold standard et l'ECMT. Par exemple, le gold standard peut considérer le libellé *douleur* tandis que l'ECMT fournira le libellé *douleurs*. Enfin, comme aucun traitement spécifique n'a été effectué pour extraire des entités qui se chevauchent, les entités imbriquées ne sont pas identifiées. L'ECMT identifie alors des concepts plus précis que ceux présents dans le gold standard, mais ces concepts ne doivent pas être considérés comme du bruit.

Les résultats obtenus pour le corpus EMEA sont proches de zéro. Ceci peut s'expliquer par la présence de caractères spéciaux dans ces textes comme « μ » qui ne sont pas traités correctement par l'ECMT. De plus, les multiples sauts de ligne présents

TABLEAU 5.3 – Résultats obtenus lors de la compétition CLEF eHealth 2015 avec l'ECMT - QUAERO Phase 1 (MEDLINE) - Reconnaissance d'entités.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT	0,2284	0,1335	0,1685	0,7091	0,6366	0,6709
Moyenne	0,3549	0,4974	0,3958	0,5232	0,7240	0,5755
Médiane	0,3878	0,5937	0,4537	0,5872	0,7897	0,6655

TABLEAU 5.4 – Résultats obtenus lors de la compétition CLEF eHealth 2015 avec l'ECMT - QUAERO Phase 1 (MEDLINE) - Reconnaissance d'entités normalisées.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT	0,2953	0,1861	0,2283	0,5003	0,3638	0,4213
Moyenne	0,3213	0,4238	0,3363	0,4280	0,5052	0,4523
Médiane	0,2953	0,4033	0,2283	0,5003	0,5735	0,4213

dans ces documents provoquent un décalage des indices ce qui entraîne les résultats constatés en correspondance exacte car l'ECMT traite chaque saut de ligne comme un espace.

5.3.4 Résultats CLEF e-Health 2016

La seconde évaluation de l'ECMT a été effectuée lors de la compétition CLEF eHealth 2016 dans la tâche d'extraction d'entités cliniques dans des textes cliniques.

Pour chaque jeu de documents MEDLINE et EMEA, le service Web ECMT a été exécuté avec les options suivantes : `restriction`, `catégorisation`, `réseau sémantique`, `priorisation` activées. Deux séries ont été réalisées : (i) la série 1 utilise 13 terminologies et ontologies : ATC, CCAM, International Classification for Patient Safety (ICPC-2), FMA, HPO, CIM10, Medline Plus, MeSH, NCIt, OMIM, HPO, Racines des médicaments (PHARMA), SNOMED Int. et (ii) la série 2 utilise 7 terminologies et ontologies : ATC, CCAM, CIM10, Medline Plus, MeSH, PHARMA, SNOMED Int..

Les résultats obtenus lors de cette compétition sont présentés dans les tableaux 5.5, 5.6, 5.7, 5.8 pour la phase 1 de reconnaissances d'entités et de reconnaissance d'entités normalisées, et dans les tableaux 5.9 et 5.10 pour la phase 2 de normalisation.

Phase 1 : Reconnaissance d'entités simples et reconnaissance d'entités normalisées Les résultats obtenus pour la phase 1 sont plutôt satisfaisants, en particulier dans la reconnaissance d'entités simples avec les résultats suivants : en correspondance exacte, nous obtenons une précision de 0,5381 et un rappel de 0,3784 (série 1) avec le corpus EMEA et une précision de 0,6407 et un rappel de 0,4375 (série 2) avec le corpus

TABLEAU 5.5 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 1 (EMEA) - Reconnaissance d'entités.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,5381	0,3784	0,4443	0,6490	0,4869	0,5564
ECMT - S2	0,5998	0,3285	0,4245	0,7175	0,4118	0,5233
Moyenne	0,5250	0,4114	0,4350	0,6377	0,5141	0,5423
Médiane	0,5998	0,3784	0,4443	0,7175	0,4808	0,5564

TABLEAU 5.6 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 1 (EMEA) - Reconnaissance d'entités normalisées.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,3800	0,2687	0,3148	0,4005	0,2842	0,3324
ECMT - S2	0,3885	0,2120	0,2743	0,4132	0,2270	0,2930
Moyenne	0,4762	0,3215	0,3761	0,4968	0,4341	0,4405
Médiane	0,4466	0,2687	0,3148	0,4666	0,2842	0,3324

MEDLINE. En correspondance inexacte, nous obtenons une précision de 0,649 et un rappel de 0,4869 (série 1) avec le corpus EMEA et une précision de 0,7668 et un rappel de 0,5865 (série 2) avec le corpus MEDLINE.

Pour la reconnaissance d'entités normalisées, en correspondance exacte, nous obtenons une précision de 0,38 et un rappel de 0,2687 (série 1) avec le corpus EMEA et une précision de 0,4776 et un rappel de 0,3271 (série 2) avec le corpus MEDLINE. En correspondance inexacte, nous obtenons une précision de 0,4005 et un rappel de 0,2842 (série 1) avec le corpus EMEA et une précision de 0,4974 et un rappel de 0,3412 (série 2) avec le corpus MEDLINE.

Phase 2 : Normalisation Les résultats obtenus pour la phase 2 sont également plutôt satisfaisants. Les résultats obtenus sont les suivants : dans l'évaluation en correspondance exacte, nous obtenons une précision de 0,6044 et un rappel de 0,4626 (série 2) avec le corpus EMEA et une précision de 0,5936 et un rappel de 0,515 (série 1) avec le corpus MEDLINE. En correspondance inexacte, nous obtenons une précision de 0,605 et un rappel de 0,463 (série 2) avec le corpus EMEA et une précision de 0,5938 et un rappel de 0,5153 (série 1) avec le corpus MEDLINE.

Dans cette phase, comme dans la phase 1, la plupart des erreurs dans les CUI identifiés sont dues à des différences entre nos données et le gold standard. Comme nous avons utilisé jusqu'à 13 terminologies de diverses sources et que le portail HeTOP ne permet pas le suivi de versions, la plupart de ces erreurs sont liées aux sources de données et peuvent également être liées aux alignements entre ces sources (et leurs

TABLEAU 5.7 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 1 (MEDLINE) - Reconnaissance d'entités.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,5399	0,4758	0,5058	0,6580	0,6492	0,6536
ECMT - S2	0,6407	0,4375	0,5199	0,7668	0,5865	0,6646
Moyenne	0,5030	0,4264	0,4455	0,6387	0,5707	0,5859
Médiane	0,6166	0,4375	0,4981	0,7394	0,5682	0,6422

TABLEAU 5.8 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 1 (MEDLINE) - Reconnaissance d'entités normalisées.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,4024	0,3562	0,3779	0,4203	0,3719	0,3946
ECMT - S2	0,4776	0,3271	0,3883	0,4974	0,3412	0,4047
Moyenne	0,5006	0,3760	0,4287	0,5181	0,4757	0,4917
Médiane	0,4927	0,3826	0,4308	0,5060	0,3917	0,4416

différentes versions) et l'UMLS.

5.3.5 Discussion

Par rapport aux résultats obtenus en 2015, les résultats obtenus en 2016 sont améliorés, en particulier dans la reconnaissance d'entités simples. Pour le jeu MEDLINE, la précision en reconnaissance exacte des entités a été améliorée de 280 % et le rappel est amélioré par plus de trois fois (voir TABLEAU 5.11). Le traitement des caractères spéciaux dans les documents et les décalages ayant été corrigés, les documents EMEA ont pu être traités en correspondance exacte et améliorés en correspondance inexacte puisque la mesure F1 était de 0,35390 en 2015 et 0,5564 (série 1) et 0,5233 (série 2) en 2016 (voir TABLEAU 5.12).

L'indexation avec plusieurs terminologies conduit à des termes dupliqués dans les

TABLEAU 5.9 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 2 (EMEA) - Normalisation.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,5669	0,4753	0,5170	0,5674	0,4757	0,5175
ECMT - S2	0,6044	0,4626	0,5240	0,605	0,4630	0,5246
Moyenne	0,5507	0,4729	0,5073	0,5511	0,4732	0,5077
Médiane	0,5669	0,4753	0,5170	0,5674	0,4757	0,5175

TABLEAU 5.10 – Résultats obtenus lors de la compétition CLEF eHealth 2016 avec l'ECMT - QUAERO Phase 2 (MEDLINE) - Normalisation.

	exact match			inexact match		
	Précision	Rappel	F1	Précision	Rappel	F1
ECMT - S1	0,5936	0,5150	0,5515	0,5938	0,5153	0,5518
ECMT - S2	0,5972	0,4676	0,5245	0,5975	0,4682	0,5250
Moyenne	0,5551	0,4854	0,5167	0,5553	0,4857	0,5170
Médiane	0,5936	0,4736	0,5245	0,5938	0,4736	0,5250

TABLEAU 5.11 – Comparatif des résultats CLEF eHealth 2015 et 2016 (série 2) obtenus avec l'ECMT dans la tâche de reconnaissance d'entités sur le corpus MEDLINE.

Tâche	Précision	Rappel	F1
2015			
entités, exact match	0,22840	0,13350	0,16850
entités, inexact match	0,70910	0,63660	0,67090
entités normalisées, exact match	0,29530	0,18610	0,22830
entités normalisées, inexact match	0,50030	0,36380	0,42130
2016			
entités, exact match	0,6407	0,4375	0,5199
entités, inexact match	0,7668	0,5865	0,6646
entités normalisées, exact match	0,4776	0,3271	0,3883
entités normalisées, inexact match	0,4974	0,3412	0,4047

résultats d'indexation qui diminuent la précision. Ce fait explique les différences qui peuvent être observées entre les séries 1 (13 terminologies) et 2 (sept terminologies). Par rapport à l'année dernière, cette question a été prise en compte et une nouvelle option a été ajoutée dans l'ECMT. Cette option `priorisation` permet de conserver uniquement les termes les plus pertinents lorsque plusieurs terminologies ajoutent un même terme dans la sortie et donc réduisent le bruit. Ce classement repose sur l'utilisation des types sémantiques. Pour chaque type sémantique, une liste des terminologies les plus pertinentes à retenir a été conçue manuellement. Cependant, à la date de la compétition, seulement 29 types sémantiques sur plus de 128 sont pris en charge. Le bruit introduit en utilisant des terminologies multiples pourrait alors être encore plus réduit dans le futur.

En revanche, certaines erreurs dans les résultats de correspondance exacte (par rapport aux résultats de correspondance inexacts) subsistent. Les erreurs introduites par l'utilisation de ressources non UMLS restent présentes bien que non quantifiées.

TABLEAU 5.12 – Comparatif des résultats CLEF eHealth 2015 et 2016 (série 2) obtenus avec l'ECMT dans la tâche de reconnaissance d'entités sur le corpus EMEA.

Tâche	Précision	Rappel	F1
2015			
entités, exact match	0.00400	0.00220	0.00280
entités, inexact match	0.43450	0.29860	0.35390
entités normalisées, exact match	0.00440	0.00240	0.00310
entités normalisées, inexact match	0.23050	0.14400	0.17730
2016			
entités, exact match	0,5998	0,3285	0,4245
entités, inexact match	0,7175	0,4118	0,5233
entités normalisées, exact match	0,3885	0,2120	0,2743
entités normalisées, inexact match	0,4132	0,2270	0,2930

5.4 Évaluation de la couverture terminologique au sein du corpus LiSSa

Plusieurs outils d'indexation sont disponibles en langue anglaise comme il a été décrit dans la section 2.5.2 ainsi que des terminologies et ontologies de référence dans le domaine de la santé, notamment l'UMLS. Les textes francophones ne bénéficient pas de ces divers outils et ressources. Le français est légèrement représenté dans l'UMLS [NÉVÉOL et al., 2014]. Dans la version 2016AA, le thésaurus UMLS français gère 9 ressources, tandis que 128 ressources sont disponibles en anglais, fournissant un concept français pour 85 685 CUI. Seulement 3.11 % des termes UMLS en anglais sont disponibles en français et, si chaque terme en anglais a une moyenne de 2 synonymes, seulement 1,54 synonymes sont disponibles pour chaque terme en français. En janvier 2017, HeTOP dispose de 363 936 CUI en français grâce à des traductions locales ou nationales. L'objectif de la présente évaluation est d'analyser la performance de ces ressources en français sur un corpus d'articles médicaux. Ceci devrait aider à réduire (i) le nombre de terminologies utilisées dans l'indexation automatique, (ii) le bruit généré en utilisant plusieurs terminologies, en particulier avec certains types spécifiques de concepts et (iii) le nombre de concepts redondants. Dans la première phase de cette étude, on analyse la couverture de 32 terminologies disponibles dans le portail HeTOP sur le corpus médical de la base de données bibliographique LiSSa [GRIFFON et al., 2014]. Ensuite, dans la deuxième phase de cette étude, les résultats d'indexation automatique des cinq principales terminologies sont évalués par rapport à une norme aurifère annotée manuellement afin d'évaluer l'indexation automatique et d'évaluer plus précisément la performance de ces cinq terminologies.

5.4.1 Le corpus LiSSa

Le corpus de la base de données bibliographiques LiSSa [9] vise à agréger l'ensemble de la littérature médicale en français. Les données de PubMed, d'Elsevier-Masson et de la revue Exercer sont disponibles, permettant la mise à disposition d'une base de données bibliographique riche de 832 446 références. Concernant les données postérieures à 2000, LiSSa regroupe 265 195 références, dont 81 239 avec le résumé en français et 209 610 avec un lien vers le texte intégral (dont 15 838 en accès gratuit). LiSSa dispose d'outils de filtre et d'export. Afin de déterminer la couverture terminologique de ce corpus, 50000 articles ont été choisis au hasard et chaque titre, résumé et ensemble de mots clés d'auteur ont été indexés à l'aide de l'outil ECMT.

5.4.2 Création du gold standard

Un jeu de données de 300 documents a été développé avec 100 titres, 100 résumés et 100 ensembles de mots-clés d'auteur. Les documents ont été choisis au hasard dans le corpus de LiSSa puis je les ai revus manuellement pour éviter les données non pertinentes (titres vides, tronqués, etc.).

5.4.3 Annotation manuelle

L'objectif de la tâche d'annotation manuelle était de fournir une ressource aussi complète que possible. Pour compléter cette tâche, un outil Web intuitif a été développé pour supporter la réalisation d'annotations structurées riches. Il inclut un support spécifique pour l'édition d'annotations : l'outil peut suggérer des concepts avec une fonction de complétion automatique répondant à l'entrée donnée. La vérification des contraintes liée à la terminologie d'indexation est entièrement intégrée dans l'interface d'annotation et les retours d'information sont immédiats, avec des effets visuels clair mettant en évidence les annotations.

Le gold standard a été annoté manuellement par quatre spécialistes (BT, CL, LS, GK) et un médecin (NG) avec les cinq terminologies obtenant la meilleure couverture du corpus déterminées lors de la première phase de l'évaluation. Cette tâche demande qu'une méthodologie spécifique soit utilisée pour réduire le biais inter-expert. Ici, les annotations ont été effectuées selon le point de vue d'un documentaliste, c'est-à-dire en annotant uniquement les principaux concepts pertinents selon l'information contenue dans le texte.

La Figure 5.5 montre l'outil d'annotation avec le texte original fourni aux annotateurs avec des annotations mises en évidence et les annotations produites.

9. http://www.lissa.fr

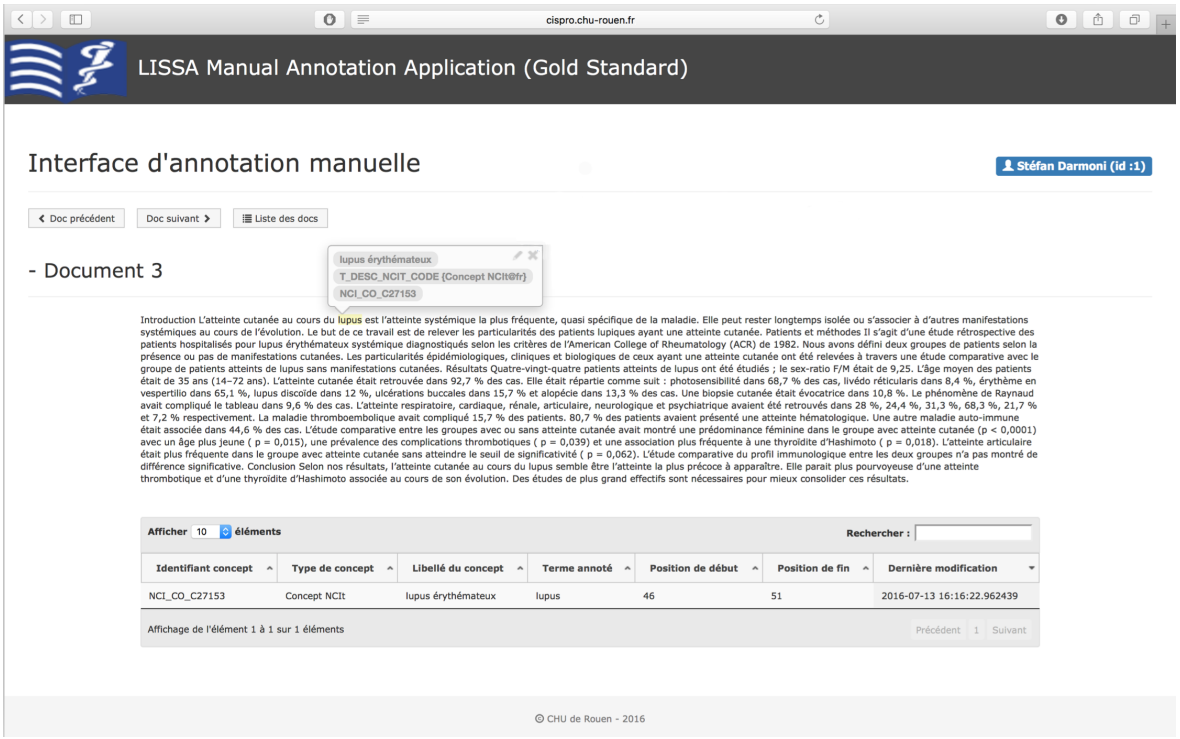

FIGURE 5.5 – Interface de l'outil d'annotation manuelle LiSSa.

5.4.4 Évaluation

Outil d'évaluation

Un outil d'évaluation dédié a été développé pour calculer les comparaisons par paires entre les annotations manuelles du gold standard et les annotations issues de l'indexation automatique. Deux concepts sont considérés comme correspondants (vrais positifs) lorsque les concepts extraits sont identiques et que leurs positions dans le texte original coïncident. L'outil calcule les vrais positifs, les faux négatifs et les faux positifs, ainsi que la précision, le rappel et la mesure F1 comme décrit dans la section 2.7.1.

Évaluation de la couverture terminologique

Dans cette tâche, 32 terminologies ont été sélectionnées parmi les 69 disponibles étant jugées les plus pertinentes pour cette tâche en tant que terminologies de référence utilisées au niveau national ou international. Ces ressources sont décrites dans le tableau A.1 en annexe. La langue source de ces ressources varie : 13 terminologies sont publiées à l'origine en français alors que 19 ont été totalement ou partiellement traduites. Pour cette évaluation, un client ECMT spécifique a été développé pour gérer la quantité de données à analyser. Chaque concept identifié dans un document et ses métadonnées (le type de concept, l'identifiant original, la terminologie) a été stocké pour une analyse ultérieure. Préalablement à l'analyse de la couverture, les termes d'indexation qui présentaient les fréquences d'occurrence les plus élevées dans tout le

corpus ont été revus manuellement pour détecter les erreurs d'indexation fréquentes. Ils ont été exclus de l'indexation dans les cas pertinents.

Le nombre de toutes les occurrences des concepts identifiées dans chaque terminologie est déterminé pour chaque catégorie de document : titres, résumés et ensembles de mots-clés. Les résultats sont détaillés dans le tableau 5.13 et la FIGURE 5.6. Des concepts distincts (c'est-à-dire comptés une seule fois) identifiés dans chaque terminologie ont également été déterminés pour chaque catégorie de documents. Les résultats sont détaillés dans le tableau 5.14.

Les cinq terminologies dont sont issus le plus de termes d'indexation dans chaque catégorie de document, NCIt, SNOMED CT, SNOMED Int., MeSH et Thésaurus Santé Publique (TSP) sont toujours les mêmes pour chaque groupe. Le thésaurus NCIt obtient la meilleure couverture dans toutes les catégories de documents, tandis que le thésaurus français de santé publique TSP est la seule ressource française à apparaître dans le premier tiers du classement des ressources. Des ressources plus spécialisées telles que Human Rare Diseases Ontology (HRDO) (maladies rares) ou ATC (thérapeutique chimique) obtiennent une couverture plus faible qu'attendu, ces ressources proposant un vocabulaire important bien que plus spécialisé. Les cinq premières terminologies donnant la meilleure couverture du corpus représentent ainsi 65 % à 70% de l'ensemble des termes d'indexation identifiés. Cependant, certaines ressources plus restreintes publiées en français réalisent une bonne couverture du corpus malgré un nombre de termes d'indexation en français limités. Ces ressources telles que le thésaurus CISMeF [DOUYÈRE et al., 2004] ou la classification Q-Codes [JAMOULLE, 2013] sont en fait développées spécifiquement pour les informations cliniques et non cliniques dans les notes cliniques et pour compléter des terminologies plus importantes telles que MeSH ou SNOMED CT.

La couverture déterminée uniquement pour des concepts distincts souligne une récurrence élevée dans toutes les terminologies et toutes les catégories de documents, en particulier dans les résumés, par nature plus longs. Pour les titres, chaque concept a une fréquence moyenne de 11,92. Pour les résumés, chaque concept a une fréquence moyenne de 75,76. Pour les mots-clés, chaque concept a une fréquence moyenne de 10,62.

Évaluation des performances

Le gold standard a été indexé avec l'ECMT avec chacune des cinq terminologies les mieux classées : MeSH, NCIt, TSP, SNOMED CT et MedDRA. SNOMED Int. a été écarté au profit de MedDRA car elle est maintenant incluse dans SNOMED CT. Les options d'indexation automatique ont été définies pour récupérer les concepts les plus larges afin de minimiser les concepts de haut niveau et de correspondre à la méthodologie adoptée lors de l'annotation manuelle. Cela permet de récupérer unique-

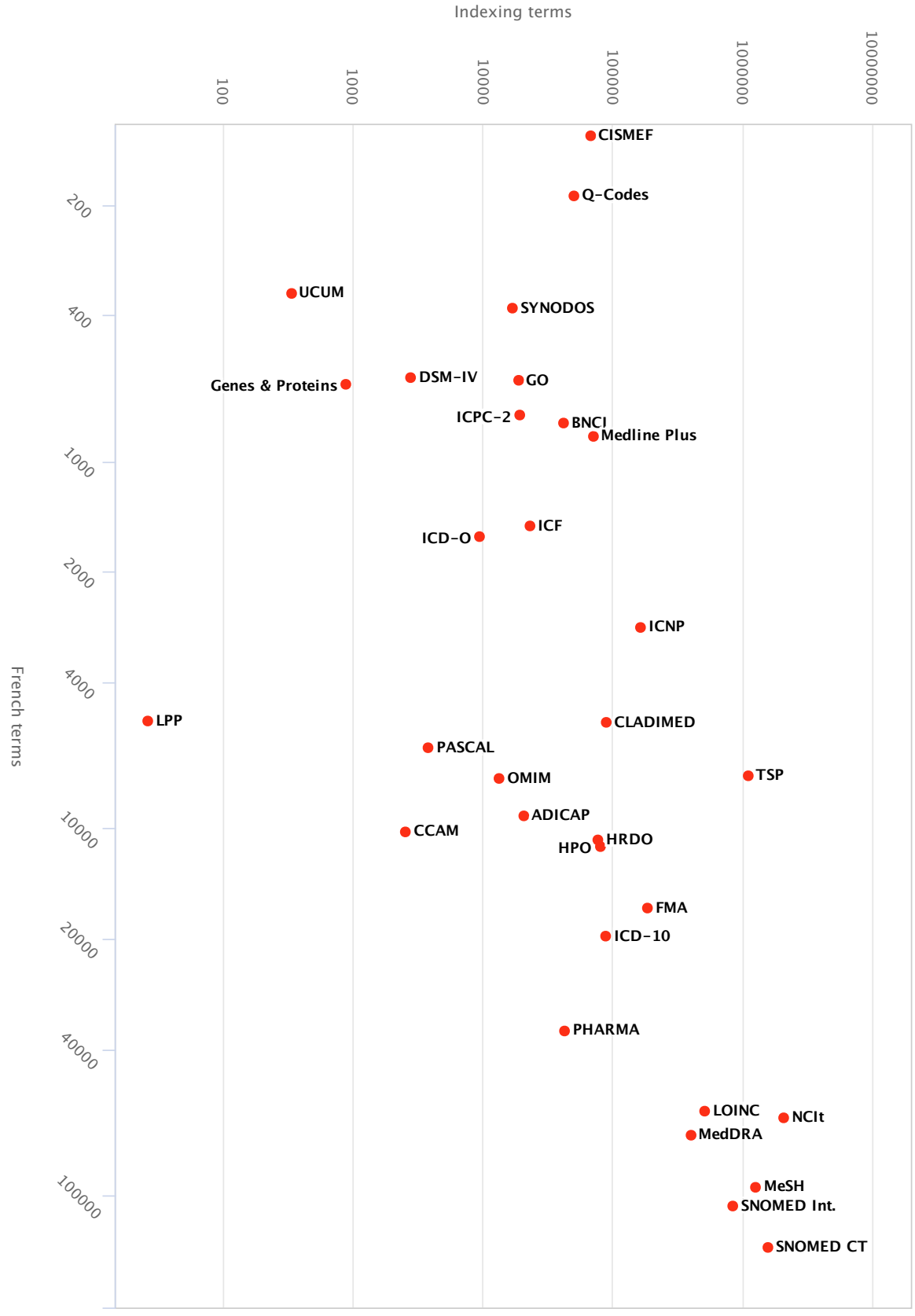

FIGURE 5.6 – Représentation graphique de la couverture terminologique des 32 terminologies et ontologies analysées dans le corpus LiSSa.

ment le concept le plus précis afin d'éviter d'introduire du bruit. Par exemple, le texte « Hypertension cardiaque » ne sera annoté qu'avec le concept *hypertension artérielle*, à l'exclusion du concept de *maladie cardiaque* plus large. Ensuite, l'outil d'évaluation a été exécuté sur les résultats d'indexation automatique et les annotations manuelles. La performance a été évaluée pour chaque terminologie et chaque catégorie de documents. Les résultats sont présentés dans le tableau 5.15.

Dans les titres, MeSH réalise les meilleures performances avec une précision de 43,49 % et un rappel de 54,93 %. MedDRA est en seconde position avec une précision de 73,47 % et un rappel de 33,33 %. TSP donne une précision de 36,64 % et un rappel de 31,48 %. SNOMED CT et NCIt atteignent une performance beaucoup plus faible avec moins de 30% de précision et un rappel comparable à ceux de TSP et de MedDRA. Selon l'analyse de couverture par concept distincts, chaque concept NCIt a une fréquence moyenne de 20,73 et SNOMED CT 11,93 dans les titres. Ces résultats attestent que NCIt et SNOMED CT contiennent de nombreux concepts redondants et non spécifiques qui introduisent du bruit dans l'indexation automatique.

Dans les mots-clés, la performance est nettement meilleure pour chaque terminologie, car les mots-clés contiennent, par définition, beaucoup moins d'informations non spécifiques que les titres et surtout les résumés. MeSH réalise les meilleures performances avec une précision de 66.80 % et un rappel de 55.71 %. MedDRA suit avec une précision de 78,89 % et un rappel de 39,66 %. TSP, NCIt et SNOMED CT atteignent des performances similaires avec une mesure F de 45±2 %.

Dans les résumés, l'indexation automatique donne de mauvais résultats avec chaque terminologie. Seul MeSH et MedDRA atteignent une mesure F1 supérieure à 20 %. Pour chaque terminologie, le rappel est similaire aux autres catégories de documents. Cependant, la précision accuse une chute très importante, en particulier NCIt et SNOMED CT donnant moins de 10 % de précision. Cela doit être comparé aux résultats de la couverture lors de la phase 1 montrant une fréquence de concept moyenne de 75,76 dans les résumés. Ces résultats valident que les résumés d'indexation entraînent un niveau élevé de bruit par rapport aux titres et aux mots clés.

TABLEAU 5.13 – Couverture terminologique dans le corpus LiSSa pour chaque catégorie de documents : titres, résumés, mots-clés.

Titres		Résumés		Mots-clés	
Terminologie	Concepts	Terminologie	Concepts	Terminologie	Concepts
NCIt	150 224	NCIt	2 040 356	NCIt	52 423
MeSH	106 170	SNOMED CT	1 543 456	MeSH	50 937
SNOMED Int.	96 771	MeSH	1 238 133	TSP	47 237
SNOMED CT	95 409	TSP	1 089 331	SNOMED Int.	45 879
TSP	84 989	SNOMED Int.	827 714	SNOMED CT	36 771
MedDRA	45 164	LOINC	502 964	MedDRA	25 802
LOINC	36 483	MedDRA	395 434	LOINC	15 491
FMA	24 300	FMA	182 398	ICNP	13 585
ICNP	24 244	ICNP	161 341	FMA	8 819
ICD-10	14 022	CLADIMED	87 924	HPO	7 633
HPO	10 493	ICD-10	86 977	Medline Plus	7 260
ATC	9 903	HPO	79 312	ICD-10	6 133
HRDO	9 612	HRDO	75 779	BNCI	3 746
CISMeF	7 961	Medline Plus	70 071	HRDO	3 553
Medline Plus	7 300	CISMeF	66 754	CISMeF	3 288
BNCI	5 141	PHARMA	49 571	CLADIMED	2 886
CLADIMED	4 934	Q-Codes	41 975	ATC	2 815
PHARMA	3 976	BNCI	41 245	ICF	2 485
ADICAP	3 875	ICF	22 757	GO	1 877
Q-Codes	3 166	ADICAP	20 408	ICPC-2	1 331
GO	2 720	ICPC-2	18 979	ADICAP	1 254
ICPC-2	2 079	GO	18 624	Q-Codes	1 191
ICF	1 820	SYNODOS	16 665	PHARMA	1 052
ICD-O	1 784	OMIM	13 147	PASCAL	721
SYNODOS	1 221	ICD-O	9 299	ICD-O	381
PASCAL	744	PASCAL	3 735	SYNODOS	331
CCAM	282	DSM-IV	2 740	DSM-IV	151
DSM-IV	176	CCAM	2 499	CCAM	98
UCUM	50	Genes & Proteins	870	UCUM	18
OMIM	36	UCUM	333	OMIM	12
LPP	0	LPP	26	LPP	0
Genes & Proteins	0	ATC	0	Genes & Proteins	0
Total	755 049	Total	8 710 817	Total	345 160

TABLEAU 5.14 – Couverture terminologique par concepts distincts dans le corpus LiSSa pour chaque catégorie de documents : titres, résumés, mots-clés.

Titres		Résumés		Mots-clés	
Terminologie	Concepts	Terminologie	Concepts	Terminologie	Concepts
MeSH	11 324	MedDRA	19 281	MeSH	5 908
SNOMED Int.	10 396	SNOMED Int.	17 968	SNOMED Int.	4 290
MedDRA	8 644	MeSH	17 353	NCIt	4 249
SNOMED CT	7 996	SNOMED CT	17 192	MedDRA	4 024
NCIt	7 247	NCIt	12 683	SNOMED CT	3 491
TSP	3 617	TSP	5 799	TSP	2 801
LOINC	1 822	FMA	3 903	LOINC	1 012
FMA	1 815	LOINC	3 372	FMA	935
ICD-10	1 758	HPO	2 997	ICD-10	870
HPO	1 488	ICD-10	2 812	HPO	823
ATC	1 332	HRDO	2 622	ICNP	712
HRDO	1 196	Q-Codes	2 488	HRDO Plus	640
ICNP	944	ADICAP	1 069	ATC	562
PHARMA	835	CLADIMED	773	PHARMA	378
ADICAP	652	Medline Plus	683	Medline Plus	352
Medline Plus	443	OMIM	669	ADICAP	272
BNCI	353	ICNP	519	BNCI	228
CLADIMED	339	BNCI	484	CLADIMED	225
ICD-O	190	ICF	415	ICPC-2	114
ICF	181	ICD-O	337	CIF	101
GO	168	GO	284	GO	97
ICPC-2	163	ICPC-2	276	ICD-O	94
CISMeF	98	SYNODOS	208	CISMeF	89
SYNODOS	72	CCAM	192	Q-Codes	84
Q-Codes	66	PHARMA	162	SYNODOS	41
CCAM	61	DSM-IV	138	DSM-IV	32
DSM-IV	38	Genes & Proteins	122	CCAM	27
PASCAL	34	CISMeF	119	Pascal	24
OMIM	19	PASCAL	44	OMIM	7
UCUM	2	UCUM	5	UCUM	3
Genes & Proteins	0	LPP	3	PHARMA	0
LPP	0	ATC	0	Genes & Proteins	0
Total	63 293	Total	114 982	Total	32 485

TABLEAU 5.15 – Résultats de l'évaluation de l'indexation automatique réalisée par l'ECMT contre le gold standard.

Terminologie	Précision	Rappel	F1
Résumés			
MeSH	12,47 %	**59,30** %	20,61 %
NCIt	5,77 %	41,99 %	10,15 %
TSP	11,08 %	37,56 %	17,11 %
MedDRA	**20,93** %	32,58 %	**25,49** %
SNOMED CT	6,24 %	32,26 %	10,45 %
Titres			
MeSH	43,49 %	**54,93** %	4**8,55** %
NCIt	25,53 %	34,04 %	29,17 %
TSP	36,64 %	31,48 %	33,86 %
MedDRA	**73,47** %	33,33 %	45,86 %
SNOMED CT	26,30 %	27,95 %	27,10 %
Mots-clés			
MeSH	66,80 %	**55,71** %	**60,75** %
NCIt	59,81 %	34,53 %	43,78 %
TSP	72,11 %	31,36 %	43,71 %
MedDRA	**78,89** %	39,66 %	52,79 %
SNOMED	62,62 %	38,17 %	47,43 %
Ensemble des résumés, titres et mots-clés			
MeSH	19,45%	**57,19**%	29,03%
NCIt	9,44%	38,08%	15,13%
TSP	16,97%	34,73%	22,80%
MedDRA	**37,06**%	35,24%	**36,13**%
SNOMED CT	10,84%	33,00%	16,32%

5.5 Synthèse

L'indexation multi-terminologique de documents biomédicaux présente des problématiques particulières traitées dans ce chapitre. Nous avons tout d'abord présenté l'outil ECMT qui permet d'identifier dans un texte des termes dans de multiples terminologies. L'utilisation de plusieurs terminologies spécifiques à un domaine pour l'indexation d'un texte médical présente des avantages. Elle peut permettre une meilleure identification de termes spécialisés comme des actes thérapeutiques ou des spécialités pharmaceutiques. De plus, la sélection de terminologies correspondant à une problématique précise plutôt que l'utilisation d'une terminologie plus généraliste permet de cibler efficacement les termes d'intérêt. Par exemple, pour l'indexation d'un corpus dans le but d'identifier des effets indésirables de médicaments, le choix de travailler avec les terminologies CIM10, Adverse Reactions Terminology (WHO-ART) et Racines des médicaments ou l'ATC serait intéressant. La multi-terminologie présente également l'inconvénient important d'augmenter considérablement le bruit, proportionnellement au nombre de ressources utilisées. Le nombre de termes identiques identifiés dans des terminologies différentes peut ainsi devenir considérable. Nous avons donc proposé une méthode permettant de prioriser l'identification d'un terme dans la terminologie la plus pertinente en fonction de son type sémantique. Cette méthode permet de réduire considérablement le nombre de termes identifiés sans affecter le rappel du système.

Parallèlement, l'outil ECMT a été évalué à plusieurs reprises dans le cadre des campagnes CLEF eHealth 2015 et 2016. Notre première participation en 2015, préalablement au développement de la méthode de priorisation, a permis de dresser un premier état des lieux et d'identifier plus précisément les points à travailler. Notre participation en 2016 a démontré les progrès réalisés avec une amélioration importante de la mesure F1 du système, en particulier pour la reconnaissance exacte. Dans un second temps, une large étude de la couverture terminologique au sein d'un corpus d'articles biomédicaux en français issus de la base bibliographique LiSSa a été réalisée. Ce projet a permis d'une part de quantifier l'apport de chaque terminologie dans l'indexation du corpus et d'autre part de conduire une évaluation de l'ECMT contre un gold standard annoté manuellement par notre équipe. Cette dernière évaluation a permis de conduire une évaluation multi-terminologique dans le sens où chaque terminologie choisie a servi à l'annotation, les campagnes de test CLEF eHealth portant elles uniquement sur les terminologies UMLS. Cette évaluation a confirmé l'intérêt du MeSH mais aussi de thésaurus comme TSP.

Malgré le travail réalisé, plusieurs aspects peuvent encore être améliorés dans le processus d'indexation. En particulier, s'agissant de l'indexation de textes médicaux, le traitement des erreurs du langage naturel est un point indispensable à traiter. En effet, l'ECMT réalise une reconnaissance exacte entre le texte et les termes. Ainsi, un texte mal orthographié, porteur d'un nombre conséquent d'abréviations, ce qui est

l'une des caractéristiques des textes comme les comptes-rendus ou les courriers, ne sera pas indexé de façon satisfaisante. Cette problématique fait l'objet du chapitre suivant.

Chapitre 6

Indexation de textes libres dans les documents médicaux

Sommaire

Dans ce chapitre, je présenterai en premier lieu les problématiques liées à l'indexation de textes médicaux narratifs comme des comptes-rendus. Nous verrons par la suite les sources et la description des données ayant été utilisées pour la conception d'une méthode d'indexation gérant le langage naturel et ses erreurs. Enfin, nous verrons l'évaluation de cette méthode dans les campagnes de test CLEF eHealth 2016 et 2017.

6.1 Indexation appliquée aux textes libres médicaux

6.1.1 L'indexation de textes médicaux narratifs

La reconnaissance d'entités a été largement étudiée au cours de la dernière décennie dans le domaine biomédical ainsi que d'autres tels que les médias sociaux [Derczynski et al., 2015] ou la reconnaissance vocale [Ma et al., 2016]. À mesure que l'utilisation des services d'indexation s'est développée, des algorithmes de pointe ont amélioré la reconnaissance d'entités dans le texte médical formel pour l'anglais [Mork et al., 2017]. Cependant, les algorithmes ont du mal à s'adapter au texte libre, car ils sont conçus pour des textes formels et sont basés sur des fonctionnalités présentes dans des textes bien formés tels que des articles biomédicaux. Dans de nombreuses applications informatiques impliquant l'enregistrement et le traitement de données personnelles, il est nécessaire de permettre des variations dans l'orthographe des mots, causées par exemple par des erreurs de transcription. Le texte libre dans les comptes-rendus ou courriers médicaux comprend des erreurs d'orthographe, une utilisation incorrecte de la ponctuation, de la grammaire et de la capitalisation [Lai et al., 2015]. Dans d'autres langues, le texte libre peut également présenter une utilisation incorrecte des marques diacritiques. Dans les rapports médicaux, le texte est généralement composé de phrases courtes ou incomplètes, semblables à la prise de notes, avec une utilisation substantielle d'abréviations ambiguës. Habituellement, les comptes-rendus cliniques sont créés avec précipitation sans relecture laissant place à un grand nombre d'erreurs. Ces erreurs ne doivent pas seulement être reliées à la complexité de la langue, mais aussi aux caractéristiques du domaine médical. Siklósi et al. ont constaté que les types d'erreurs les plus fréquentes sont les erreurs d'orthographe, les erreurs grammaticales, les phrases tronquées et les abréviations non standardisées [Siklósi et al., 2016]. En effet, par opposition au texte formel, les abréviations sont rarement définies dans les rapports médicaux. Malgré les efforts déployés dans l'indexation, même dans le domaine biomédical, l'extraction de l'information dans les notes cliniques doit encore faire face à ces défis Menasalvas et Gonzalo-Martin [2016]. Pour traiter cette problématique et plus particulièrement celle de la variabilité du texte libre, il apparaît nécessaire d'utiliser des méthodes du traitement automatique de la langue liées à la reconnaissance

partielle de texte, qui peuvent être combinées à des méthodes statistiques décrites dans la section 2.5.2.

6.1.2 La reconnaissance partielle de texte : mesures de similarité

Les mesures de similarité de texte jouent un rôle important dans la recherche et les applications liées aux textes dans des tâches telles que la RI, la classification des textes, le regroupement de documents, la détection des thèmes, les systèmes de questions/réponses, la traduction automatique, la synthèse de textes et d'autres. La similarité entre les mots est une partie fondamentale de la similarité du texte qui peut ensuite être utilisée pour déterminer des similarité de phrases, de paragraphes ou de documents. Les mots peuvent être similaires au niveau lexical ou au niveau sémantique. Lexicalement, les mots sont similaires s'ils ont une séquence de caractères similaire. Sémantiquement, les mots sont semblables s'ils ont le même sens, sont utilisés de la même manière ou dans le même contexte. La similarité lexicale est introduite par différents algorithmes basés sur les chaînes de caractères, la similarité sémantique est introduite à travers des algorithmes basés sur le corpus ou les connaissances. Les mesures basées sur les chaînes de caractères fonctionnent sur des séquences de chaînes et la composition des caractères. Une mesure de similarité entre deux chaînes de caractères est une mesure appréciant la similarité ou la dissemblance (distance) entre deux chaînes de texte pour une correspondance ou une comparaison approximative des chaînes. La similarité basée sur le corpus est une mesure de similarité sémantique qui détermine la similarité entre les mots en fonction de l'information obtenue à partir de grands corpus. La similarité fondée sur la connaissance est une mesure de similarité sémantique qui détermine le degré de similarité entre les mots en utilisant des informations dérivées des réseaux sémantiques.

Mesures de similarité lexicales

Les mesures de similarité lexicales se basent sur les séquences de chaînes de caractères et leur composition en caractères. Une métrique de chaîne est une mesure qui détermine la similarité ou la dissemblance (distance) entre deux chaînes de caractères pour une correspondance ou une comparaison approximative des chaînes.

Mesures basées sur les caractères

Longest Common SubString (LCS) L'algorithme de la sous-chaîne commune la plus longue considère que la similarité entre deux chaînes est basée sur la longueur de la chaîne contiguë de caractères qui existe dans les deux chaînes.

Damerau-Levenshtein Damerau-Levenshtein définit la distance entre deux chaînes en comptant le nombre minimal d'opérations nécessaires pour transformer une chaîne en l'autre, où une opération est définie comme une insertion, une suppression ou une substitution d'un seul caractère ou une transposition de deux caractères adjacents [Hall et Dowling, 1980; Levenshtein, 1966; Peterson, 1980].

Jaro Jaro est basé sur le nombre et l'ordre des caractères communs entre deux chaînes. Il prend en compte les écarts d'orthographe typiques [Jaro, 1995, 2012].

Jaro-Winkler Jaro-Winkler est une extension de la distance de Jaro. Elle utilise une échelle de préfixe qui donne des valeurs plus favorables aux chaînes qui correspondent dès le début à une longueur de préfixe définie [Winkler, 1990].

Needleman-Wunsch L'algorithme Needleman-Wunsch est un exemple de programmation dynamique et a été la première application de programmation dynamique à la comparaison des séquences biologiques. Il effectue un alignement global pour trouver le meilleur alignement sur l'ensemble des deux séquences. Il est approprié lorsque les deux séquences sont de longueur similaire, avec un degré significatif de similarité [Needleman et Wunsch, 1970].

Smith-Waterman Smith-Waterman est un autre exemple de programmation dynamique. Cet algorithme effectue un alignement local pour trouver le meilleur alignement sur le domaine conservé de deux séquences. Il est utile pour des séquences dissemblables qui sont suspectées de contenir des régions de similarité ou des motifs de séquence similaires dans leur contexte de séquence plus large [Smith et Waterman, 1981].

N-gram N-gram est une sous-séquence de n éléments à partir d'une séquence de texte donnée. Les algorithmes de similarité N-gram comparent les n-gram de chaque caractère ou mot de deux chaînes de caractères. La distance est calculée en divisant le nombre de n-gram similaires par le nombre maximal de n-gram [Barrón-Cedeño et al., 2010].

Mesures basées sur les termes

Distance de bloc La distance de bloc est également connue sous le nom de distance de Manhattan, la distance de valeur absolue ou la distance L1. Cet algorithme calcule la distance qui serait parcourue pour passer d'un point de données à l'autre si un chemin de grille est suivi. La distance de bloc entre deux éléments est la somme des différences de leurs composants correspondants [Krause, 1973].

Similarité des cosinus La similarité des cosinus est une mesure de la similarité entre deux vecteurs d'un espace de produit interne qui mesure le cosinus de l'angle entre eux.

Coefficient de Dice Le coefficient de Dice est défini comme le double du nombre de termes communs dans les chaînes comparées divisé par le nombre total de termes dans les deux chaînes [Dice, 1945].

Distance euclidienne La distance euclidienne ou la distance L2 est la racine carrée de la somme des différences carrées entre les éléments correspondants des deux vecteurs.

Distance de Jaccard La similarité de Jaccard est calculée comme le nombre de termes partagés sur le nombre de tous les termes uniques dans les deux chaînes [Jaccard, 1901].

Coefficient de correspondance Le coefficient de correspondance est une approche basée sur un vecteur très simple qui compte simplement le nombre de termes et dimensions similaires, sur lesquels les deux vecteurs sont non nuls.

Coefficient de chevauchement Le coefficient de chevauchement est similaire au coefficient de Dice, mais considère deux chaînes comme correspondantes si l'une est un sous-ensemble de l'autre.

Mesures de similarité sémantiques basées sur le corpus

La similarité basée sur le corpus est une mesure de similarité sémantique qui détermine la similarité entre les mots en fonction de l'information obtenue à partir de grands corpus. Un corpus est une grande collection de textes écrits ou parus qui est utilisée pour la recherche linguistique.

Hyperspace Analogue to Language (HAL) Hyperspace Analogue to Language (HAL) [Lund et Burgess, 1996; Lund et al., 1995] crée un espace sémantique à partir de cooccurrences de mots. Une matrice mot par mot est formée avec chaque élément matriciel étant la force d'association entre le mot représenté par la ligne et le mot représenté par la colonne. L'utilisateur de l'algorithme a alors la possibilité d'abandonner les colonnes d'entropie faible de la matrice. Au fur et à mesure que le texte est analysé, un mot-clé est placé au début d'une fenêtre de dix mots qui enregistre les mots voisins comptés comme cooccurrents. Les valeurs de matrice sont accumulées en pondérant la cooccurrence inversement proportionnelle à la distance par rapport au mot-clé. On pense que les mots proches voisins reflètent davantage la sémantique du

mot-clé et sont donc plus élevés. HAL enregistre également l'information de l'ordre des mots en traitant la cooccurrence différemment selon que le mot voisin est apparu avant ou après le mot-clé.

L'analyse sémantique latente (LSA) L'analyse sémantique latente (LSA) [Landauer et Dumais, 1997] est la technique de similarité basée sur le corpus la plus populaire. LSA suppose que les mots qui ont une signification proche se produiront dans des textes similaires. Une matrice contenant des nombres de mots par paragraphe (les lignes représentent des mots uniques et des colonnes représentent chaque paragraphe) est construite à partir d'un grand texte et une technique mathématique appelée décomposition de valeur singulière (SVD) est utilisée pour réduire le nombre de colonnes tout en préservant la structure de similarité entre les lignes. Les mots sont ensuite comparés en prenant le cosinus de l'angle entre les deux vecteurs formés par deux lignes quelconques.

L'analyse sémantique latente généralisée (GLSA) L'analyse sémantique latente généralisée (GLSA) [Matveeva et al., 2005] étend l'approche LSA en se concentrant sur les vecteurs de termes au lieu de la double représentation terme/document. GLSA requiert une mesure de l'association sémantique entre les termes et une méthode de réduction de la dimension. L'approche GLSA peut combiner toute mesure de similarité sur l'espace des termes avec n'importe quelle méthode appropriée de réduction de la dimension. La matrice traditionnelle document/terme est utilisée dans la dernière étape pour fournir les poids dans la combinaison linéaire des vecteurs de termes.

L'analyse sémantique explicite (ESA) L'analyse sémantique explicite (ESA) [Gabrilovich et Markovitch, 2007] est une mesure utilisée pour calculer la relation sémantique entre deux textes arbitraires. La technique basée sur Wikipedia représente les termes (ou les textes) en tant que vecteurs à haute dimension. Chaque entrée vectorielle présente le poids TF-IDF entre le terme et un article Wikipédia. La relation sémantique entre deux termes (ou textes) est exprimée par la mesure du cosinus entre les vecteurs correspondants.

Cross-Language Explicit Semantic Analysis (CL-ESA) CL-ESA [Potthast et al., 2008] est une généralisation multilingue de l'ESA. CL-ESA exploite une collection de référence multilingue alignée sur le document telle que Wikipedia pour représenter un document en tant que vecteur concept indépendant de la langue. La relation de deux documents dans différentes langues est évaluée par la similarité du cosinus entre les représentations vectorielles correspondantes.

Google Distance La distance Google Distance (NGD) normalisée [Cilibrasi et Vitanyi, 2007] est une mesure de similarité sémantique dérivée du nombre de résultats renvoyés par le moteur de recherche Google pour un ensemble donné de mots-clés. Les mots-clés ayant les mêmes significations ou similaires en langage naturel ont tendance à être proches dans des unités de Google distance, alors que les mots ayant des significations différentes ont tendance à être plus éloignés. Plus précisément, la distance NGD entre deux termes de recherche x et y est :

$$NGD(x,y) = \frac{max\left(\log f(x), \log f(y)\right) - \log f(x,y)}{\log M - min\left(\log f(x), \log f(y)\right)} \tag{6.1}$$

où M représente le nombre total de pages Web recherchées par Google ; $f(x)$ et $f(y)$ sont le nombre de résultats pour les termes de recherche x et y, respectivement ; et $f(x,y)$ est le nombre de pages Web sur lesquelles x et y se produisent. Si les deux termes de recherche x et y ne se produisent jamais ensemble sur la même page Web, mais se produisent séparément, la distance NGD standard entre eux est infinie. Si les deux termes se produisent toujours ensemble, leur NGD est nul ou équivalent au coefficient entre x^2 et y^2.

Combinaison et comparaison des approches

Une méthode pour mesurer la similarité sémantique entre les phrases ou les textes très courts, basée sur l'information sémantique et l'ordre des mots, a été présentée dans Li et al. [2006]. Tout d'abord, la similarité sémantique découle d'une base de connaissances lexicale et d'un corpus. Deuxièmement, la méthode proposée considère l'impact de l'ordre des mots sur la signification de la phrase. La similarité de l'ordre des mots mesure le nombre de mots différents ainsi que le nombre de paires de mots dans un ordre différent. Les auteurs de Islam et Inkpen [2008] ont présenté une méthode nommée Semantic Text Similarity (STS). Cette méthode détermine la similarité de deux textes à partir d'une combinaison d'informations sémantiques et syntaxiques. Ils ont considéré deux fonctions obligatoires (similarité de chaîne et distance sémantique) et une fonction facultative (similarité d'ordre des mots communs). La méthode STS a obtenu un très bon coefficient de corrélation de Pearson pour les séries de données de 30 phrases et a surpassé les résultats obtenus dans Li et al. [2006]. Les auteurs de Aggarwal et al. [2012] ont présenté une approche qui combine la mesure de la relation sémantique basée sur le corpus sur toute la phrase avec les scores de similarité sémantique basés sur la connaissance qui ont été obtenus pour les mots relevant des mêmes rôles syntaxiques dans les deux phrases. Tous les scores en tant que caractéristiques ont été alimentés par des modèles d'apprentissage, comme la régression linéaire, et des modèles de sacs de mots pour obtenir un seul score donnant le degré de similarité entre les phrases. Cette approche a montré une amélioration significative dans le calcul de la

similarité sémantique entre les phrases en combinant la mesure de similarité fondée sur la connaissance et la mesure de similarité basée sur le corpus contre la mesure basée sur le corpus prise seule. Une corrélation prometteuse entre les résultats de similarité manuelle et automatique a été obtenue dans Buscaldi et al. [2012] en combinant deux modules. Le premier module calcule la similarité entre les phrases en utilisant la similarité n-gram, et le second module calcule la similarité entre les concepts dans les deux phrases en utilisant une mesure de similarité de concept et WordNet. Un système nommé UKP avec des résultats de corrélation raisonnables a été introduit dans Bär et al. [2012], il a utilisé un modèle simple de régression linéaire basé sur des données d'apprentissage, pour combiner plusieurs mesures de similarité de texte. Ces mesures étaient la similarité des chaînes de caractères, la similarité sémantique, les mécanismes d'expansion du texte et les mesures liées à la structure et au style. Les modèles finaux de l'UKP se composaient d'une combinaison log-linéaire d'environ 20 caractéristiques, sur les 300 caractéristiques possibles mises en œuvre.

6.1.3 L'approche phonétique

La correspondance phonétique est utilisée pour identifier les chaînes qui peuvent être de prononciation similaire, quelle que soit leur orthographe. Une application typique est l'identification de noms propres. Par exemple, un opérateur téléphonique reçoit verbalement un nom, en suppose l'orthographe (ou il lui est fourni une orthographe, qui peut être incorrecte), et utilise son hypothèse pour interroger une base de données de noms. Le système de correspondance phonétique doit alors trouver dans la base de données les chaînes les plus susceptibles d'être identiques ou similaires à celles de la requête. Comme il n'existe aucun moyen fiable de déterminer automatiquement la prononciation d'une chaîne, cette correspondance doit être inexacte. La plupart des algorithmes existants et qui sont décrits ci-après ont été conçus pour la langue anglaise. Par conséquent, l'application de leurs règles aux mots d'autres langues peut ne pas donner un résultat exploitable. Un certain nombre d'algorithmes ont été développés pour l'appariement des mots, c'est-à-dire qui tentent d'identifier les variations d'orthographe du mot, dont l'un des plus connus est l'algorithme de Soundex. À l'heure actuelle, il existe un certain nombre d'algorithmes phonétiques utilisant différents degrés de complexité pour surmonter ces variations. La sélection suivante de techniques illustre la gamme actuelle d'approches du problème sous leur forme plus générale, car beaucoup de ces algorithmes peuvent et ont été modifiés pour produire des correspondances plus précises pour des applications plus spécialisées.

Russell Soundex L'algorithme Soundex est conçu principalement pour être utilisé avec des noms en anglais. Il a été breveté par Robert C. Russell et Margaret King Odell

en 1918 [1] et 1922 [2]. L'algorithme convertit chaque nom en un code à quatre caractères, qui peut être utilisé pour identifier des noms équivalents, et qui est déterminé comme suit :

1. Conserver la première lettre du nom et enlever toutes les occurrences de a, e, h, i, o, u, w, y dans d'autres positions.
2. Affecter les chiffres suivants aux lettres restantes après la première :
   — b, f, p, v ← 1
   — l ← 4
   — c, g, j, k, q, s, x, z ← 2
   — m, n ← 5
   — d, t ← 3
   — r ← 6
3. Si deux ou plusieurs lettres avec le même code étaient adjacentes dans le nom d'origine (avant l'étape 1), les omettre tous sauf le premier
4. Convertir en forme « lettre, chiffre, chiffre, chiffre » en ajoutant des zéros à gauche (s'il y a moins de trois chiffres), ou en enlevant les chiffres les plus à droite, s'il y en a plus de trois.

Par exemple, les noms propres Euler, Gauss, Hilbert, Knuth et Lloyd reçoivent les codes respectifs E460, G200, H416, K530, L300. Bien que le Soundex soit plus précis que d'utiliser uniquement des similarités de caractères entre les noms, l'algorithme n'est pas idéal. Par exemple, « reynold » et « renauld » sont tous deux réduits à *r543*, mais, plus communément, Soundex fait l'erreur de transformer des chaînes de sonorité dissemblables telles que « catherine » et « cotroneo » vers le même code et de transformer des chaînes de sons similaires en différents codes. Il n'y a pas de classement des correspondance : les chaînes sont similaires ou non similaires. Pour les noms communs, il se révèlera plus efficace pour reconnaître le début d'un mot, par exemple dans des fonction d'auto-complétion.

Metaphone Cette technique a été développée par Lawrence Philips pour faire correspondre des mots qui se ressemblent et se basent sur les règles classiques de la prononciation anglaise. Metaphone ignore les voyelles après la première lettre et réduit l'alphabet restant à seize sons consonnes, bien que les voyelles soient conservées lorsqu'elles sont la première lettre. Les lettres en double ne sont pas ajoutées au code. Zero est utilisé pour représenter le « th », et « X » est utilisé pour le son « sh ». Les seize sons de consonnes sont : B X S K J T F H L M N P R Ø W Y. Il est plus précis que

1. US patent 1261167, R. C. Russell, 1918-04-02
2. US patent 1435663, R. C. Russell, 1922-11-14

le soundex car il prend en compte les règles de base de la prononciation anglaise. Metaphone est disponible comme en standard dans de nombreux systèmes. L'algorithme produit des clés en sortie. Les sonorités similaires des mots partagent les mêmes clés et sont de longueurs différentes.

Double Metaphone Le Double Metaphone est la deuxième génération de l'algorithme Metaphone [Philips, 2000]. Il est conçu principalement pour coder les noms anglais américains tout en tenant compte du fait que de tels mots peuvent avoir plus d'une prononciation acceptable. Double Metaphone peut calculer un encodage primaire et secondaire pour un mot ou un nom donné pour indiquer à la fois la prononciation la plus probable ainsi qu'une prononciation alternative optionnelle (d'où le « double » dans le nom). DM essaie de tenir compte d'une multitude d'irrégularités en anglais, mais aussi dans les langues slave, germanique, celtique, grecque, française, italienne, espagnole, chinoise et d'autres langues. Bien que puissant, DM a ses limites et ses inconvénients. DM a été conçu pour rechercher des listes de noms propres plutôt que de grandes quantités de texte. L'encodage peut ne pas correspondre à des mots mal orthographiés qui modifient sérieusement la structure phonétique du mot. Malgré ses limites, l'algorithme de DM, qui est libre, est toujours un système d'encodage phonétique flexible et puissant aujourd'hui, en particulier dans une approche multilingue.

Caverphone 2.0 L'algorithme de correspondance phonétique Caverphone [Hood, 2004] a été créé par David Hood dans le projet Caversham à l'Université d'Otago en Nouvelle-Zélande en 2002, révisé dans une version 2 en 2004. Il a été créé pour aider à la correspondance des données dans des listes électorales établies entre la fin du 19ème siècle et le début du 20ème siècle, où le nom ne devait être que sous une « forme communément reconnaissable ». L'algorithme est destiné à s'appliquer à ces noms qui ne pouvaient pas être facilement comparables entre les listes électorales, après que les correspondances exactes ont été retirées du groupe de correspondances potentielles. Caverphone 2.0 considère plus efficacement les cas présentant des voyelles initiales différentes et tolère mieux les consonnes post-vocaliques que Métaphone. La sensibilité de Caverphone 2.0 aux faux positifs se situe entre Metaphone et Soundex (tendant à être similaire à Metaphone).

Parmi ces algorithmes, l'algorithme DM est donc un choix intéressant pour un système multilingue puisqu'il permet la prise en compte de multiples langages. Il a donc été sélectionné pour la suite de mes travaux.

6.2 Sources de données

Dans le cadre de la campagne CLEF eHealth 2016, un des deux tâches proposées impliquait l'indexation de certificats de décès, rédigés à la main par des médecins, avec la CIM10. Ses données décrites dans cette partie ont servi de base à la conception d'une méthode prenant en compte les irrégularités dans les textes médicaux rédigés à la main.

6.2.1 Le corpus français CépiDC

Depuis 1968, le laboratoire CépiDC de l'Institut National de la Santé Et de la Recherche Médicale (INSERM), se consacre à l'élaboration annuelle des statistiques nationales sur les causes des décès en association avec l'Institut national français de la statistique et des études économiques (Insee), la diffusion des données et les études et recherches sur les causes médicales du décès. Ces statistiques sont construites à partir d'informations provenant de certificats de décès. L'équipe CépiDC gère une base de données contenant plus de 18 000 000 dossiers de décès Pavillon et Laurent [2003]. À partir de ces données, des jeux de données ont été conçus dans le cadre des challenges CLEF eHealth 2016 et 2017. Ce challenge se concentre sur la reconnaissance de l'entité nommée dans les certificats de décès en français et en anglais (en 2017). Alors que les certificats de décès sont des documents normalisés remplis par les médecins pour signaler la mort d'un patient, ils présentent habituellement des erreurs d'orthographe ou de frappe, des abréviations et, en français, un texte non accentué ou un mélange de casse. Dans l'édition 2016, un jeu de données a été utilisé dans le but d'indexer les certificats de décès à l'aide de la CIM10. Dans l'édition 2017, deux jeux de données ont été fournis. La tâche de ce challenge consiste à extraire les codes CIM10 des textes bruts du certificat de décès. Cette tâche d'extraction d'information repose sur le texte fourni pour extraire les codes CIM10 des certificats, ligne par ligne.

Corpus CépiDC - CLEF eHealth 2017

Deux jeux de données sont fournis. Le premier ensemble de données s'appelle « dataset aligné » et le second s'appelle « dataset brut ».

Dataset aligné L'ensemble de données comprend 31 690 certificats de décès traités par CépiDC en 2014 totalisant 91 962 lignes. Les annotations dans le corpus CépiDC se composent de codes CIM10 et ont été attribuées par ligne de texte. L'ensemble de données est fourni dans un fichier CSV. Chaque ligne contient douze champs d'information associés à une ligne brute de texte à partir d'un certificat de décès d'origine comme suit :

— DocID : l'identifiant du certificat de décès ;

— YearCoded : année de traitement du certificat de décès ;

— Gender : genre de la personne décédée ;

— Age : âge au moment de la mort, arrondi ;

— LocationOfDeath : lieu du décès ;

— LineID : numéro de la ligne au sein du certificat de décès ;

— RawText : text brut saisi dans le certificat de décès ;

— IntType : intervalle de temps durant lequel le patient a souffert de la cause codée, selon les catégories suivantes : minutes, heures, jours, mois, années ;

— IntValue : intervalle de temps durant lequel le patient a souffert de la cause codée ;

— CauseRank : rang du code CIM10 ;

— StandardText : entrée du dictionnaire ou extrait du texte auant permis de déterminer le code CIM10 ;

— ICD10 : code CIM10 code associé à la ligne de texte.

La sortie comprend les 9 champs de saisie plus deux champs de texte (CauseRank et StandardText) utilisés pour indiquer la preuve supportant le code CIM10 fourni dans le douzième et dernier champ.

Dataset brut Les données de 31 683 certificats de décès sont distribuées parmi trois fichiers CSV. Le premier fichier inclut les champs suivants : DocID, YearCoded, LineID, RawText, IntType, IntValue. Le deuxième fichier inclut les champs suivants : DocID, YearCoded, Gender, PrimCauseCode, Age, LocationOfDeath. Le troisième fichier inclut les champs suivants : DocID, YearCoded, LineID.

Corpus CépiDC - CLEF eHealth 2016

L'ensemble de données comprend 65 843 certificats de décès traités par CépiDC au cours de la période 2006-2012. Le corpus est fourni au format CSV et chaque ligne contient douze champs d'information associés à une ligne brute de texte à partir d'un certificat de décès d'origine. La sortie comprend les 9 champs de saisie plus deux champs de texte utilisés pour signaler les éléments de preuve mettant en évidence le code CIM10 fourni dans le douzième et dernier champ. Le dixième champ doit contenir l'extrait du texte original ayant permis la prédiction du code CIM10.

6.2.2 Le corpus anglais CDC

Ce corpus a été utilisé dans le cadre du challenge CLEF eHealth 2017 et mis au point par l'institut CDC américain. Les données de 6 665 certificats de décès sont distribuées parmi trois fichiers CSV. Le premier fichier inclut les champs suivants :

DocID, YearCoded, LineID, RawText, IntType, IntValue. Le deuxième fichier inclut les champs suivants : DocID, YearCoded, Gender, PrimCauseCode, Age, LocationOfDeath. Le troisième fichier inclut les champs suivants : DocID, YearCoded, LineID.

La Figure 6.1 donne un exemple de documents fournis dans le corpus français CépiDC. Le septième champ contient le texte à annoter, le onzième l'entrée du dictionnaire ICD10 correspondant au texte et au dernier champ le code ICD10 correspondant. De même, la Figure 6.2 donne un exemple de documents fournis dans le corpus anglais CDC.

```
64185;2013;2;85;2;5;SYNDROME DE GLISEMENT AVEC GRABATISATION DEPUIS
OCTOBRE 2012;4;3;6-1;syndrome glissement;R453
64185;2013;2;85;2;5;SYNDROME DE GLISEMENT AVEC GRABATISATION DEPUIS
OCTOBRE 2012;4;3;6-1;grabatisation;R263
79317;2013;2;85;2;6;héùorragie digestive basse sur surdosage en
AVK;3;5;6-3;hémorragie digestive basse;K921
79317;2013;2;85;2;6;héùorragie digestive basse sur surdosage en
AVK;3;5;6-3;surdosage avk;X44
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;laryngectomie totale;Z900
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;abcès cervical;L021
64370;2013;1;80;2;5;ABCES CERVICAL . LARYNGECTOMIE TOTALE.ATCD
D'IDM.;NULL;NULL;;antécédent infarctus myocarde;I258
```

FIGURE 6.1 – Extraits de certificats de décès en français dans le corpus CépiDC.

```
13496;2015;;;;6;Senile dementia of Alzheimer's type ASHD;;;;senile
dementia;F03
13496;2015;;;;6;Senile dementia of Alzheimer's type
ASHD;;;;alzheimer;G309
13496;2015;;;;6;Senile dementia of Alzheimer's type ASHD;;;;ashd;I251
16915;2015;;;;2;HEALTHCAREASSOCIATED PNEUMONIA;;;;healthcare-associated
pneumonia;J189
```

FIGURE 6.2 – Extraits de certificats de décès en anglais dans le corpus CDC.

6.2.3 Dictionnaires

Dictionnaires orthographiques

Le corpus français de CépiDC comprend six versions d'un dictionnaire CIM10 organisé manuellement, développé à CépiDC correspondant aux années : 2006-2010, 2011, 2012, 2013, 2014 et 2015. En 2017, le corpus CDC anglais comprend un dictionnaire

CIM10 organisé manuellement, développé par la CDC fournissant 170 285 entrées. Ces ressources ont été utilisées pour créer des dictionnaires d'orthographe. De plus, les données d'apprentissage ont été utilisées pour compléter ces dictionnaires.

Traitement des dictionnaires

Pour chaque langue, les versions du dictionnaire ont été fusionnées si nécessaire. Chaque terme de la CIM10 a été séparé en mots et les doublons ont été supprimés. Les deux listes de mots uniques obtenues ont fourni un dictionnaire d'orthographe pour chaque langue.

Ensuite, un dictionnaire supplémentaire a été produit à partir de chaque jeu de données d'apprentissage en extrayant le code CIM10 et les combinaisons de termes. Le nombre de fois qu'un code CIM10 a été utilisé dans le corpus de formation a également été déterminé. Pour des termes ambigus, c'est-à-dire des termes qui correspondent à plus d'un code CIM10, le terme le plus utilisé a été conservé. Chaque dictionnaire supplémentaire a été fusionné avec les dictionnaires fournis dans le corpus correspondant. Si un terme était présent dans le dictionnaire additionnel et un dictionnaire de corpus, mais les codes correspondants étaient différents, le code du dictionnaire supplémentaire a été supprimé pour éviter d'introduire une ambiguïté entre les versions du dictionnaire. Ce traitement a permis de compléter les dictionnaires fournis en particulier par quelques abréviations manquantes.

6.3 Extraction d'information dans des textes libres médicaux à l'aide de la CIM-10 : CIM-IND

CIM-IND est conçu pour faire correspondre les termes CIM10 au texte comme entrée dans la version pertinente de l'CIM10. L'extraction est effectuée au niveau de la phrase du texte en utilisant des techniques de traitement du langage naturel. Le système est construit en utilisant des extensions Python et Python/C et fournit une réponse au format CSV pour chaque concept identifié avec : (i) le texte d'entrée, (ii) le décalage du premier et du dernier mot contenu dans le concept de santé, (iii) l'identifiant CIM10 et (iv) le terme CIM10. CIM-IND effectue trois étapes principales pour identifier les termes de la CIM10 : une étape de prétraitement, la sélection des candidats et le classement des candidats.

6.3.1 Pré-traitements

Normalisation

Plusieurs étapes de prétraitement sont effectuées, y compris le filtrage des mots d'arrêt (en utilisant les listes de mots d'arrêt NLTK par défaut pour le français et l'anglais Bird et al. [2009]) et le filtrage élision (en supprimant les articles abrégés qui sont contractés avec des termes). Les mots sont insensibles à la casse. Les diacritiques dans les textes français sont conservés et Unicode est utilisé pour correspondre.

Correction orthographique

La vérification orthographique est effectuée avec la bibliothèque Enchant en utilisant le dictionnaire construit manuellement comme décrit dans la section 6.2.3.

6.3.2 Sélection des candidats

Une méthode basée sur l'algorithme de codage phonétique Double Metaphone (DM) [Philips, 2000] est utilisé pour exploiter une première recherche de terme approximative.

Tout d'abord, CIM-IND calcule l'encodage DM pour chaque mot inclus dans la phrase normalisée. Ensuite les candidats dont le terme CIM10 présente un encodage DM correspondant sont récupérés. Cette étape fournit rapidement une liste de candidats pertinents de la CIM10 et permet d'effectuer des traitements coûteux sur un ensemble réduit de termes dans la phase suivante. De cette façon, le système repose sur une base de données pour stocker le codage DM précalculé pour chaque mot disponible dans chaque dictionnaire de la version CIM10. Cette base de données est implémentée avec le système MySQL et son modèle est décrit dans la figure 6.3. Il permet la gestion de versions des terminologies intégrées.

Par exemple, dans la figure 6.1, les lignes 1-2 contiennent le mot mal orthographié « glisement » (pour « glissement ») et les lignes 3-4 contiennent le mot mal orthographié « héúorragie » (pour « hémorragie »). La première erreur est correctement traitée par l'algorithme DM qui fournit le même encodage à la fois pour le mot mal orthographié et le mot correct. Cependant, la deuxième erreur n'est pas correctement traitée. Comme l'erreur modifie profondément la phonétique du mot, l'algorithme DM donne un encodage différent de celui du mot correct. Cela met en évidence l'importance de procéder à une vérification orthographique du texte normalisé pour éviter les mots les plus mal orthographiés avant l'encodage phonétique et ainsi assurer une liste de candidats correcte.

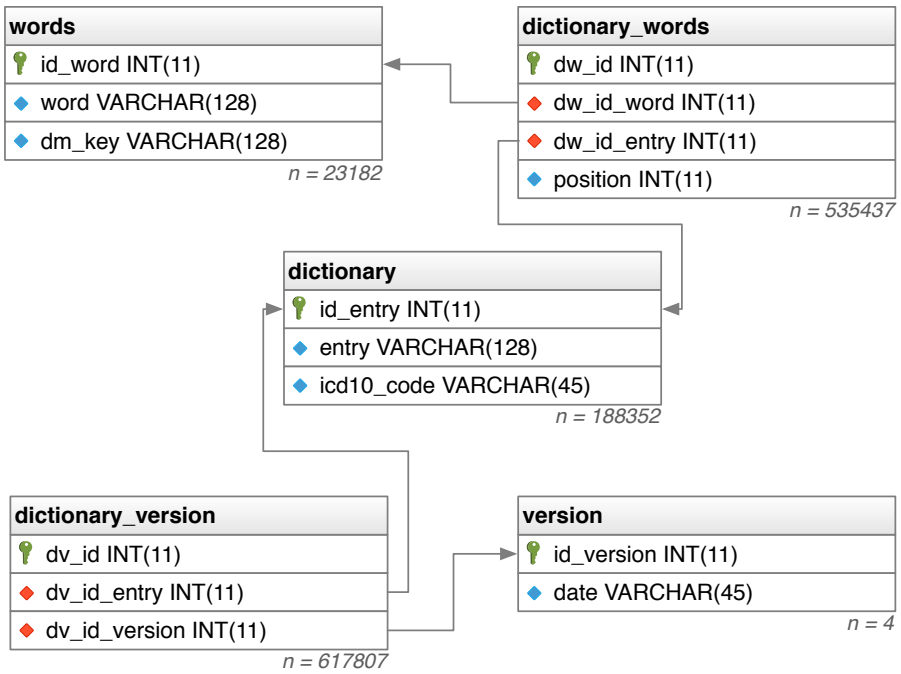

FIGURE 6.3 – Modèle de données physique du système CIM-IND.

6.3.3 Classement des candidats

Enfin, un algorithme de score de distance pondéré (WDS) a été développé pour classer la liste des termes candidats. L'algorithme WDS renvoie un score de similarité de 0 à 100 pour chaque candidat, 100 représentant une correspondance parfaite. Le terme le plus probable ayant le score le plus élevé est conservé en tant que terme correspondant à la CIM10. Comme un seul ou plusieurs termes CIM10 peuvent être présents dans une phrase, deux cas sont considérés. Tout d'abord, si la longueur de la chaîne du candidat s_1 est semblable à la longueur de la chaîne s_2 (c.-à-d. qu'un seul terme CIM10 est attendu), deux scores sont calculés : (i) un score de base (BS) et (ii) un score d'ensemble (SeS). Le BS est calculé en déterminant la distance de Levenshtein entre les séquences s_1 et s_2 échelonnée de 0 à 100. Le SeS détermine tous les caractères alphanumériques dans chaque chaîne et les traite comme un ensemble. Ensuite, deux chaînes de caractères sont construites en concaténant, d'une part, l'intersection triée et, d'autre part, le reste trié. Ensuite, la distance de ces chaînes est calculée en contrôlant les correspondances partielles non ordonnées.

Sinon, si l'une des séquences est au moins 1,5 fois plus longue que l'autre, deux scores partiels sont calculés : (i) un score de base partiel (PBS) et (ii) un score d'ensemble partiel (PSeS). Le PBS renvoie la distance de la sous-chaîne la plus similaire sous la forme d'un nombre entre 0 et 100. Tout d'abord, chaque bloc représentant une séquence de caractères correspondants dans une chaîne est déterminé. Ensuite, la meilleure reconnaissance partielle sera l'alignement contenant au moins un de ces blocs.

Le PSeS calcule le PBS pour chaque chaîne créée à partir de l'intersection triée et le reste trié de s_1 et s_2. Pour s'assurer que seuls les résultats complets peuvent retourner une correspondance parfaite, les scores partiels sont mis à l'échelle en fonction de la longueur de s_1 et s_2. Tous les scores d'ensemble sont mis à l'échelle de 0,95. Enfin, le score WDS est déterminé comme le maximum de ces scores.

6.4 Application aux corpus CépiDC et CDC

Le système CIM-IND a été évalué lors des éditions 2016 [Cabot et al., 2016b; Kelly et al., 2016; Névéol et al., 2016] et 2017 [Cabot et al., 2017; Goeuriot et al., 2017; Neveol et al., 2014b] de la compétition CLEF eHealth. Ces deux éditions ont proposé une tâche d'extraction des causes de décès dans les certificats de décès fournis par les corpus CépiDC et CDC telles que codées par la CIM-10.

6.4.1 Compétition CLEF eHealth 2016

Résultats préliminaires et comparaison des méthodes de similarité

Préalablement à l'évaluation lors de la compétition, plusieurs méthodes ont été évaluées sur le jeu d'apprentissage fourni. L'analyse du jeu d'apprentissage a été effectuée à l'aide de quatre méthodes : (i) tout d'abord avec l'ECMT, (ii) une méthode combinant correction orthographique et reconnaissance exacte du texte, (iii) une méthode combinant correction orthographique et reconnaissance partielle à l'aide de la distance de Levenshtein, et enfin (iv) une méthode combinant la correction orthographique du texte et la reconnaissance partielle déterminée par les distances de Levenshtein et l'algorithme LCS. Les résultats obtenus sont présentés dans le Tableau 6.1.

Résultats de la compétition CLEF eHealth 2016

CIM-IND analyse l'ensemble des données de test CépiDC et produit les résultats au format CSV. Les résultats tels qu'évalués par l'organisation de la compétition sont présentés dans le tableau 6.2. Dans ce jeu de données, une précision de 0,6964 et un rappel de 0,6634 sont obtenus. Le nombre de termes retrouvés est satisfaisant, mais en comparaison des résultats obtenus par d'autres équipes participant à la compétition, le taux d'erreur est élevé (Tableau 6.3). Dans cette version préliminaire de CIM-IND, seules les distances de Levenshtein et de sous-chaîne commune la plus longue étaient utilisées pour classer les candidats. Ces deux mesures aboutissent ainsi à un nombre important d'erreurs, mettant en évidence la nécessité de travailler à une combinaison de méthode de distances plus complète et adaptée avec le score WDS.

TABLEAU 6.1 – Comparaison des méthodes de similarité sur le corpus CépiDC.

Méthode	ECMT	Correction orthographique et reconnaissance exacte	Correction orthographique et distance de Levenshtein	Correction orthographique et distances de Levenshtein et LCS
Précision	0,69	0,64	0,79	0,78
Rappel	0,23	0,62	0,48	0,59
F1	0,34	0,63	0,60	0,67
Temps d'exécution par ligne	40ms	70ms	100ms	100ms

TABLEAU 6.2 – Résultats de CIM-IND sur le corpus CépiDC - CLEF eHealth 2016.

	TP	FP	FN	Précision	Rappel	F1
CIM-IND	72192	31480	36626	0,6964	0,6634	0,6795
Moyenne				0,7878	0,6636	0,7185
Médiane				0,8110	0,6554	0,6997

TABLEAU 6.3 – Résultats du challenge CLEF eHealth 2016 par équipe pour la tâche de codage CIM10 sur le corpus CépiDC (de NÉVÉOL et al. [2016]).

Équipe	Précision	Rappel	F1
Erasmus-run2	0,886	**0,813**	**0,848**
Erasmus-run1	**0,890**	0,803	0,844
ERIC-ECSTRA-run2	0,882	0,655	0,752
ERIC-ECSTRA-run1	0,811	0,615	0,700
CIM-IND	*0,696*	*0,663*	*0,680*
LIMSI-run1	0,765	0,569	0,652
BITEM-run1	0,585	0,526	0,554
Moyenne	0,788	0,664	0,719
Médiane	0,811	0,655	0,700

6.4.2 Compétition CLEF eHealth 2017

L'évaluation a été effectuée par les organisateurs pour deux cas : (i) pour tous les codes CIM10, l'évaluation principale et (ii) pour les codes CIM10 traitant d'un type particulier de décès, appelés « causes externes » ou décès violents Névéol et al. [2017].

Tableau 6.4 – Résultats de CIM-IND pour chaque jeu de données français CépiDC et anglais CDC - CLEF eHealth 2017.

	Toutes causes			Causes externes		
Jeu de données	Précision	Rappel	F1	Précision	Rappel	F1
Français - Jeu de données brut CépiDC	0,8568	0,6886	0,7636	0,5670	0,4310	0,4897
Français - Jeu de données aligné CépiDC	0,8346	0,7751	0,8038	0,5343	0,4717	0,5011
Anglais - Jeu de données CDC	0,8393	0,7827	0,8100	0,4261	0,3889	0,4066

Résultats sur le corpus français CépiDC

CIM-IND a été exécuté sur les deux jeux de test français et une exécution a été soumise pour chacun de ces jeux de données. Le Tableau 6.4 montre les résultats obtenus sur les jeux de données brut et aligné. Les résultats obtenus par chaque équipe participante dans le challenge sont présentés dans les tableaux 6.5 et 6.6.

Sur le jeu de données brutes, CIM-IND a obtenu une précision de $0,8568$ et un rappel de $0,6886$ ($F1 = 0,7636$) pour tous les codes CIM10. En ce qui concerne uniquement les codes CIM10 correspondant à des causes externes (c'est-à-dire des décès violents), CIM-IND a réalisé une performance inférieure substantielle avec une précision de $0,567$ et un rappel de $0,431$ ($F1 = 0,4897$).

Sur le jeu de données aligné, CIM-IND a atteint une précision de $0,8346$ et un rappel de $0,7751$ ($F1 = 0,8038$) pour tous les codes CIM10. En ce qui concerne uniquement les codes CIM10 correspondant à des causes externes, CIM-IND a de nouveau réalisé une performance inférieure avec une précision de $0,5343$ et un rappel de $0,4717$ ($F1 = 0,5011$).

Étant donné que la principale différence entre ces deux jeux de données était liée au formatage, il était attendu d'obtenir des résultats similaires. Cependant, remarquablement, le jeu de données alignées obtient un rappel plus élevé que le jeu de données brutes. Ensuite, il convient de noter que les performances sont considérablement inférieures en ce qui concerne uniquement les codes CIM10 liés aux causes externes pour les deux jeux de test. Dans ce corpus, les résultats de CIM-IND sont considérablement meilleurs que le score moyen et médian de toutes les séries soumises.

TABLEAU 6.5 – Résultats du challenge CLEF eHealth 2017 par équipe pour la tâche de codage CIM10 sur le jeu de données brut CépiDC français (de NÉVÉOL et al. [2017]).

		Toutes causes			Causes externes		
	Équipe	Précision	Rappel	F1	Précision	Rappel	F1
Essais officiels	CIM-IND	**0,857**	**0,689**	**0,764**	**0,567**	**0,431**	**0,490**
	LITL-run2	0,666	0,414	0,510	0,443	0,367	0,401
	LIRMM-run1	0,541	0,480	0,509	0,443	0,367	0,401
	LIRMM-run2	0,540	0,480	0,508	0,560	0,283	0,376
	LITL-run1	0,651	0,404	0,499	0,538	0,277	0,365
	TUC-MI-run2	0,044	0,026	0,033	0,010	0,004	0,005
	TUC-MI-run1	0,025	0,015	0,019	0,006	0,005	0,005
	Moyenne	0,475	0,358	0,406	0,367	0,247	0,292
	Médiane	0,541	0,414	0,508	0,443	0,283	0,376
Essais non officiels	LIMSI-run2	0,872	0,784	0,825	0,700	0,594	0,643
	LIMSI-run1	0,883	0,760	0,817	0,709	0,559	0,625
	TUC-MI-run1-corr.	0,883	0,539	0,669	0,780	0,290	0,423
	TUC-MI-run2-corr.	0,882	0,536	0,667	0,767	0,283	0,414
	UNIPD-run1	0,629	0,468	0,537	0,350	0,381	0,365
	UNIPD-run2	0,518	0,384	0,441	0,362	0,251	0,296
	Mondeca-run1	0,375	0,131	0,194	0,335	0,228	0,271

Résultats sur le corpus anglais CDC

Une série a été soumise pour le jeu de données CDC anglais. Le TABLEAU 6.4 présente les résultats obtenus sur ce jeu de données. Les résultats obtenus par chaque équipe participante dans le challenge sont présentés dans le tableau 6.7.

CIM-IND a obtenu une précision de $0,8393$ et un rappel de $0,7827$ ($F1 = 0,8100$) pour tous les codes CIM10. En ce qui concerne uniquement les codes CIM10 correspondant à des causes externes, CIM-IND a atteint une performance inférieure avec une précision de $0,4261$ et un rappel de $0,3889$ ($F1 = 0,4066$).

En ce qui concerne tous les codes CIM10, ces résultats sont légèrement meilleurs que les résultats obtenus avec le jeu de données brutes français, mais remarquablement similaires à ceux obtenus avec le jeu de données alignées. Encore une fois, il existe une baisse significative de la performance concernant uniquement les codes CIM10 liés aux causes externes. Dans ce cas, les résultats sont inférieurs à ceux obtenus sur les deux jeux de données français, à la fois pour la précision et le rappel. Dans ce corpus, selon les deux évaluations, nos résultats sont supérieurs au score moyen et médian de toutes les séries soumises lors de la compétition.

TABLEAU 6.6 – Résultats du challenge CLEF eHealth 2017 par équipe pour la tâche de codage CIM10 sur le jeu de données aligné CépiDC français (de NÉVÉOL et al. [2017]).

		Toutes causes			Causes externes		
	Équipe	Précision	Rappel	F1	Précision	Rappel	F1
Essais officiels	SIBM-run1	0,835	**0,775**	**0,804**	0,534	**0,472**	**0,501**
	WBI-run1	0,780	0,751	0,765	**0,740**	0,318	0,445
	TUC-MI-run2	**0,874**	0,611	0,719	0,412	0,403	0,407
	LITL-run1	0,612	0,550	0,579	0,412	0,403	0,407
	LIRMM-run1	0,506	0,530	0,518	0,482	0,348	0,404
	LIRMM-run2	0,505	0,530	0,517	0,534	0,275	0,363
	LITL-run2	0,646	0,402	0,495	0,709	0,151	0,249
	TUC-MI-run1	0,426	0,297	0,350	0,218	0,119	0,154
	Moyenne	0,648	0,555	0,593	0,505	0,311	0,366
	Médiane	0,629	0,540	0,548	0,508	0,333	0,406
Essais non officiels	LIMSI-run2	0,854	0,881	0,867	0,630	0,674	0,651
	LIMSI-run1	0,865	0,865	0,865	0,640	0,636	0,638
	TUC-MI-run1-corr.	0,875	0,614	0,722	0,748	0,323	0,452
	UNIPD-run1	0,604	0,517	0,557	0,320	0,402	0,356
	UNIPD-run2	0,488	0,418	0,451	0,376	0,265	0,311

6.4.3 Discussion

Le développement de CIM-IND a débuté pour faire face aux difficultés d'indexation des textes libres médicaux avec l'ECMT et explorer les différentes solutions permettant de traiter efficacement les insconsistences du texte libre. En 2016, le système a été évalué dans la tâche CLEF eHealth correspondante, uniquement sur un corpus français et a obtenu un score F1 de 0,6795, ce qui était légèrement inférieur à la moyenne des résultats CABOT et al. [2016b]. Ces premiers résultats ont démontré la nécessité d'apporter diverses améliorations concernant notamment le classement des candidats et le traitement des fautes d'orthographe et autres erreurs lexicales. Les résultats obtenus en 2017 mettant en œuvre le score WDS ont démontré ces améliorations avec une augmentation de 12% du score F1 dans le jeu de données brutes français et une augmentation de 18% du score F1 dans le jeu de données alignées français. En outre, cette deuxième évaluation a démontré que CIM-IND obtient des résultats satisfaisants à la fois en français et en anglais, avec des résultats supérieurs à la moyenne dans les deux langues.

Cependant, certains aspects de nos résultats devraient être approfondis. Bien que CIM-IND ait obtenu des résultats satisfaisants, certaines erreurs liées à la désambiguïsation ou aux fautes d'orthographe demeurent. En particulier, les fautes d'orthographe significatives sur des mots qui ne font pas partie du dictionnaire d'orthographe en-

TABLEAU 6.7 – Résultats du challenge CLEF eHealth 2017 par équipe pour la tâche de codage CIM10 sur le jeu de données CDC anglais (de NÉVÉOL et al. [2017]).

		Toutes causes			Causes externes		
	Équipe	Précision	Rappel	F1	Précision	Rappel	F1
Essais officiels	KFU-run1	0,893	0,811	**0,850**	0,584	0,357	**0,443**
	KFU-run2	0,891	**0,812**	0,850	0,631	0,325	0,429
	TUC-MI-run1	**0,940**	0,725	0,819	0,426	0,389	0,407
	CIM-IND	0,839	0,783	0,810	0,233	**0,524**	0,323
	TUC-MI-run2	0,929	0,717	0,809	0,232	**0,524**	0,322
	WBI-run1	0,616	0,606	0,611	0,880	0,175	0,291
	WBI-run2	0,616	0,606	0,611	**1,00**	0,159	0,274
	LIRMM-run1	0,691	0,514	0,589	0,168	0,262	0,205
	LIRMM-run2	0,646	0,527	0,580	0,292	0,111	0,161
	Unipd-run1	0,496	0,442	0,468	0,246	0,119	0,160
	UNSW-run1	0,401	0,352	0,375	0,246	0,119	0,160
	Unipd-run2	0,382	0,341	0,360	0,279	0,095	0,142
	UNSW-run2	0,371	0,328	0,348	0,043	0,310	0,076
	Mondeca-run1	format invalide			format invalide		
	Moyenne	0,670	0,582	0,622	0,405	0,267	0,261
	Médiane	0,646	0,606	0,611	0,279	0,262	0,274
Non officiel	LIMSI-run2	0,899	0,801	0,847	0,723	0,373	0,492
	LIMSI-run1	0,909	0,765	0,831	0,837	0,325	0,469
	Mondeca-run1	0,691	0,309	0,427	0,042	0,056	0,048

traîneraient un encodage phonétique incorrect, et donc une liste incorrecte de termes candidats. En anglais, les résultats pourraient être légèrement améliorés avec une terminologie plus complète ou un jeu d'apprentissage plus large pour couvrir certains termes manquants, en particulier les abréviations. En outre, la baisse de performance concernant les codes CIM10 liés aux causes externes devrait être étudiée et semble affecter toutes les séries soumises lors de la compétition. Les causes externes présentent un contexte spécifique et souvent une terminologie spécifique liée aux accidents, aux décès violents ou aux surdoses induites par le traitement. Ils apparaissent plus rarement dans les jeux d'apprentissage. En fait, seules 2 440 lignes dans le jeu de données français (110 869 lignes) et 313 lignes dans le le d'apprentissage anglais (39 333 lignes) semblent être liées à des causes externes (codes CIM10 V01 à Y98). Cela peut expliquer dans une certaine mesure la performance réduite. En outre, dans certains cas, les codes CIM10 associés à une ligne donnée utilisent le contexte fourni dans d'autres lignes du même certificat de décès. CIM-IND traite chaque ligne de manière indépendante et n'a pas pu annoter correctement ces lignes. Enfin, il faut noter que la combinaison de

méthodes statistiques à l'approche actuelle pourrait s'avérer une voie d'amélioration de ces résultats.

6.5 Synthèse

Dans ce chapitre nous avons vu la problématique d'indexation de textes narratifs, rédigés à la main, et la prise en compte des erreurs de langage qu'ils peuvent comporter. Pour répondre à cette problématique, j'ai conçu une méthode basée sur la reconnaissance phonétique du texte et mis au point un score de similarité permettant de sélectionner le meilleur terme candidat. Elle permet d'identifier dans un texte porteur d'erreur des termes dans une terminologie choisie. Ici, la CIM10 a été choisie dans le cadre de la compétition CLEF eHealth mais cette méthode est adaptable pour d'autres terminologies et doit être à court terme utilisée dans notre outil multi-terminologique ECMT. Cette méthode a été évaluée dans le cadre des éditions 2016 et 2017 CLEF eHealth avec de très bons résultats en français et en anglais. Par la suite, l'un des axes d'amélioration de cette méthode devrait se concentrer sur le traitement des abréviations. En effet, là encore, les textes comme les comptes-rendus posent un défi spécifique. Alors que les abréviations sont usuellement définies dans des articles médicaux, elles ne le sont que très rarement dans un compte-rendu médical où elles peuvent de plus se révéler nombreuses. La résolution de ces abréviations qui peuvent s'avérer ambiguës est donc un développement futur important.

Chapitre 7

Conclusion et perspectives

Les objectifs de cette thèse s'inscrivent ainsi dans la large problématique de recherche d'information dans les données issues du DPI. Les aspects abordés dans cette problématique ont été multiples : d'une part la mise en œuvre d'une recherche d'information clinomique au sein du DPI, à travers la modélisation de données hétérogènes et multidisciplinaires, l'intégration de données, informations et connaissances provenant de la biologie moléculaire et le développement d'outils d'interrogation et de visualistion et d'autre part la recherche d'information au sein de données non structurées issues du DPI à travers l'indexation de textes médicaux narratifs à l'aide de réseaux sémantiques.

L'un des premiers objectifs fut le recensement et l'étude des SRI disponibles en santé. On peut retenir en particulier dans cette conclusion le projet américain i2b2 [1] faisant référence dans le domaine de l'exploitation des données cliniques contenues dans les dossiers patients informatisés, ainsi que le projet dérivé tranSMART [2] dédié à la réutilisation des données omiques. Bien que partageant une partie de leur modèle de données, ces deux solutions ciblent des besoins très différents. i2b2 se concentre sur l'exploitation des données contenues dans des entrepôts de données cliniques pour la création de cohortes de patients et plus largement pour répondre à des problématiques de recherche clinique. Sa portée est donc réservée à une approche multi-patients et n'autorise pas l'exploration des données d'un patient individuel. tranSMART propose lui également une approche multi-patients et plus spécifiquement la fouille de données clinomique pour leur réutilisation en recherche fondamentale. Il s'inscrit ainsi pleinement dans une optique de recherche translationnelle. Dans la première partie de cette thèse, nous avons ainsi exploré une approche différente. L'objectif principal était ainsi la réalisation d'un système de recherche d'information intégrant données cliniques et données omiques mais également combinant à la fois une approche multi-patients, dédiée aux problématiques de recherche clinique et de santé publique, et une approche mono-patient, plus proche de la pratique clinique. La combinaison de ces deux ap-

1. http://www.i2b2.org
2. http://transmartfoundation.org

proches peut être argumentée. Un système unique permet de lever plusieurs verrous : l'unification des données médicales, cliniques, biologiques ou d'imagerie pour la transmission et la non-redondance des informations et la continuité des connaissances entre recherche et pratique clinique indispensable à la personnalisation des soins. Certains avantages peuvent être signalés en termes de formation nécessaire aux utilisateurs. La réalisation de ce système a fait intervenir plusieurs étapes. Tout d'abord le recensement et la formalisation des différents types de données omiques et la sélection des données pertinente à intégrer. Cette étape a abouti à la conception d'un modèle de données conceptuel générique des données omiques, étendant le modèle de données cliniques mis en œuvre dans le projet ANR TecSAN RAVEL. Ce modèle de données permet de gérer un grand nombre des types de données existants en biologie moléculaire : variants structuraux, variants du nombre de copies, variants SNP et insertions/délétions, données d'expression de gènes, exons, miRNA, jonctions, transcrits et protéines, données de méthylation de l'ADN. Ce modèle a été validé par l'intégration de données issues de sources diverses et hétérogènes : le portail américain The Genome Cancer Atlas et les laboratoires de recherche en oncologie INSERM U918 et U614. Cette partie a nécessité le développement d'outils logiciels assurant l'extraction, le traitement, l'intégration et la validation des données. Dans le cadre de ce travail, j'ai également été amenée à intégrer de nouvelles données dans le portail terminologique HeTOP : la base de données NCBI Gene, la base de données Uniprot/Swissprot et enfin la base de données Online Mendelian Inheritance in Man (OMIM). Ces données viennent enrichir le portail terminologique de données concernant respectivement les gènes, protéines et maladies génétiques humaines. Le moteur de recherche développé dans le cadre du projet RAVEL pour les données cliniques a été adapté au modèle de données omiques afin de permettre la recherche d'information à la fois dans les données cliniques et dans les données omiques. Enfin, une interface de visualisation des données intégrées a été développée et intégrée au prototype de Dossier Patient informatisé. Ce prototype permet la visualisation et l'interrogation des données omiques intégrées ainsi que des données cliniques issues de 2 000 dossiers patients informatisés du Centre Hospitalier Universitaire de Rouen. Ce démonstrateur exploite les données intégrées dans le portail terminologique HeTOP.

Dans la seconde partie de cette thèse, je me suis intéressée à l'indexation et la recherche d'information dans les documents médicaux tels que les comptes-rendus médicaux. Dans ce cadre, j'ai participé en 2015 au challenge CLEF E-Health, en particulier à la tâche consacrée à l'extraction d'information dans des textes cliniques. Ce challenge nous a permis d'évaluer l'outil ECMT (Extracteur de Concepts Multi-Terminologique) qui permet d'indexer des textes à partir des terminologies exploitées dans le portail terminologique HeTOP. Suite à cette première évaluation, j'ai réalisé plusieurs tâches visant à optimiser l'extraction des concepts, notamment en concevant une méthode

adressant la problématique de duplication des concepts propre à l'indexation multi-terminologique. Parallèlement, plusieurs corpus ont été analysés en collaboration avec la Cité des Sciences et l'Université Nice Sophia Antipolis. Ce travail a été évalué dans le cadre de l'édition 2016 du challenge CLEF eHealth. Ce challenge a amené le développement d'une méthode d'indexation spécifique à la classification internationale des maladies appliquée au traitement de certificats de décès. Cet outil implémente le suivi des différentes versions d'une terminologie ainsi que le traitement des erreurs du langage naturel. L'évaluation de l'indexation multi-terminologique d'un corpus d'articles biomédicaux issus de la base de données bibliographique LiSSa (http ://www.lissa.fr) a été réalisée pour évaluer l'apport des différentes terminologies disponibles en français dans le portail terminologique HeTOP. Cette évaluation s'est déroulée en deux phases, la première nous permettant d'étudier quantitativement l'apport de chaque terminologie dans l'indexation des documents, la seconde nous permettant d'évaluer qualitativement l'indexation multi-terminologique et l'apport de chaque ressource. Cette évaluation a fait appel à un gold standard annoté manuellement par l'équipe et au développement de l'outil d'annotation nécessaire à cette tâche.

Par la suite, plusieurs perspectives peuvent être envisagées pour approfondir le travail effectué pendant cette thèse. L'intégration de données cliniques et omiques et leur exploitation au sein d'un système de recherche d'information au sein du DPI sont appelées à évoluer, d'une part à travers les progrès accomplis dans les disciplines liées à la biologie moléculaire, tant d'un point de vue des connaissances fondamentales acquises que des progrès techniques, et d'autre part à travers l'évolution des ressources nécessaires à leur exploitation. Les problématiques de disponibilité, confidentialité et sécurité des données restent aujourd'hui un frein important à l'accomplissement de cet objectif et à la mise en production et la diffusion d'un tel système au sein des systèmes d'information hospitaliers. La problématique de recherche d'information dans les données non structurées telles que les comptes-rendus médicaux contenus dans les DPI peut être approfondie tant dans les possibilités offertes par la linguistique et le traitement automatisé du langage que par les approches statistiques. Les compétitions organisées régulièrement dans ce domaine ont montré une grande diversité dans les approches possibles, mais aussi une certaine maturité des outils proposés. De tels outils offrent des perspectives intéressantes dans de multiples applications à la recherche d'information. En particulier, les travaux de cette thèse autour de l'ECMT ont permis une valorisation de cet outil auprès de la société Alicante en l'intégrant dans sa suite logicielle et l'installant dans un hôpital lillois, nous a permis un retour sur de nouveaux cas d'usage qui devraient permettre la poursuite des travaux sur la recherche d'information et l'extraction de concepts. Sur cette dernière thématique, une nouvelle thèse va débuter pour tester une approche « deep learning » (bourse CIFRE avec la startup OMICX).

Liste des publications

Cabot, C., L. F. Soualmia et S. J. Darmoni. 2017 (submitted), «SIBM at CLEF eHealth Evaluation Lab 2017 : Multilingual Information Extraction with CIM-IND», dans *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2017)*. URL http://ceur-ws.org/Vol-1609/16090047.pdf.

Cabot, C., L. F. Soualmia, J. Grosjean, N. Griffon et S. J. Darmoni. 2017, «Evaluation of the Terminology Coverage in the French Corpus LiSSa», *Studies in health technology and informatics*, vol. 235, p. 126–130.

Cabot, C., L. F. Soualmia, B. Dahamna et S. J. Darmoni. 2016a, «SIBM at CLEF eHealth Evaluation Lab 2016 : Extracting Concepts in French Medical Texts with ECMT and CIMIND», dans *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2016)*, vol. 1609, p. 47–60. URL http://ceur-ws.org/Vol-1609/16090047.pdf.

Cabot, C., R. Lelong, J. Grosjean, L. F. Soualmia et S. J. Darmoni. 2016b, «Retrieving Clinical and Omic Data from Electronic Health Records», *Studies in health technology and informatics*, vol. 221, p. 115–115.

Cabot, C., L. F. Soualmia, B. Dahamna et S. J. Darmoni. 2016c, «ECMT : Indexation multi-terminologique de documents biomédicaux», dans *1er Forum Franco-Québécois d'Innovation en Santé, Polytechnique Montréal.*

Cabot, C., L. F. Soualmia et S. J. Darmoni. 2016d, «Recherche d'information dans les données cliniques et omiques au sein du Dossier Patient Informatisé», Journée Plateforme d'Indexation 2.0.

Lelong, R., C. Cabot, L. F. Soualmia et S. J. Darmoni. 2016, «Semantic Search Engine to Query into Electronic Health Records with a Multiple-Layer Query Language», dans *MEDIR workshop*. URL http://medir2016.imag.fr/data/MEDIR_2016_paper_8.pdf.

Cabot, C., L. F. Soualmia et S. J. Darmoni. 2015a, «Intégration de données cliniques et omiques pour la recherche d'information dans le Dossier Patient Infor-

matisé», dans *Actes des 26èmes Journées Francophones d'Ingénierie des Connaissances (IC), associées à la Plateforme de l'Association Française pour l'Intelligence Artificielle*, Rennes, France, p. 183–193. URL https://hal.archives-ouvertes.fr/hal-01179292/.

Cabot, C., L. F. Soualmia, J. Grosjean, R. Lelong et S. J. Darmoni. 2015b, «Integrating and Retrieving Clinical and Omic Data in Electronic Health Records», dans *7th International Workshop on Knowledge Representation for Health Care (KRH4C) and 8th International Workshop on Process-oriented Information Systems in Healthcare (ProHealth)*, p. 154–159. URL https://www.researchgate.net/profile/Lina_Soualmia/publication/280066101_Integrating_and_Retrieving_Clinical_and_Omic_Data_in_Electronic_Health_Records/links/55a7a47408aeceb8cad65695.pdf.

Soualmia, L. F., C. Cabot, B. Dahamna et S. J. Darmoni. 2015, «SIBM at CLEF e-Health Evaluation Lab 2015», dans *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2015)*, vol. 1391.

Lelong, R., C. Cabot, T. Merabti, J. Grosjean, N. Griffon, B. Dahamna, P. Massari et S. Darmoni. 2015, «Information Retrieval in Electronic Health Records Using a Multiple Layer Query Language», dans *Journées Recherche en Imagerie et Technologies pour la Santé 2015*, pp 128-129.

Cabot, C., L. F. Soualmia, J. Grosjean, R. Lelong et S. J. Darmoni. 2015, «Integrating and retrieving clinical and omic data in electronic health records», 15 ans du Master de Bioinformatique de l'Université de Rouen.

Cabot, C., J. Grosjean, R. Lelong, A. Lefebvre, T. Lecroq, L. F. Soualmia et S. J. Darmoni. 2014a, «Omic Data Modelling for Information Retrieval», dans *IWBBIO*, Citeseer, p. 415–424.

Cabot, C., M. Mary, C. Saad, A. Renaux, A. Bertrand, A. Velt, A. Lefebvre, C. Bérard, N. Vergne et H. Dauchel. 2014b, «GC- VC/DGE : a user-friendly web application for Going over Concordance across results from NGS bioinformatics analytic pipelines», dans *ECCB 2014*, Strasbourg, France.

Cabot, C., M. Mary, C. Saad, A. Renaux, A. Bertrand, A. Velt, A. Lefebvre, C. Bérard, N. Vergne et H. Dauchel. 2014c, «GC- VC/DGE : a user-friendly web application for Going over Concordance across results from NGS bioinformatics analytic pipelines», 3ème Journée Scientifique de l'IRIB.

Coutant, S., C. Cabot, A. Lefebvre, M. Léonard, E. Prieur-Gaston, D. Campion, T. Lecroq et H. Dauchel. 2012, «EVA : Exome Variation Analy-

zer, an efficient and versatile tool for filtering strategies in medical genomics», *BMC bioinformatics*, vol. 13, n ^{o} Suppl 14, p. S9.

Annexe A

Annexes

A.1 Figures annexes

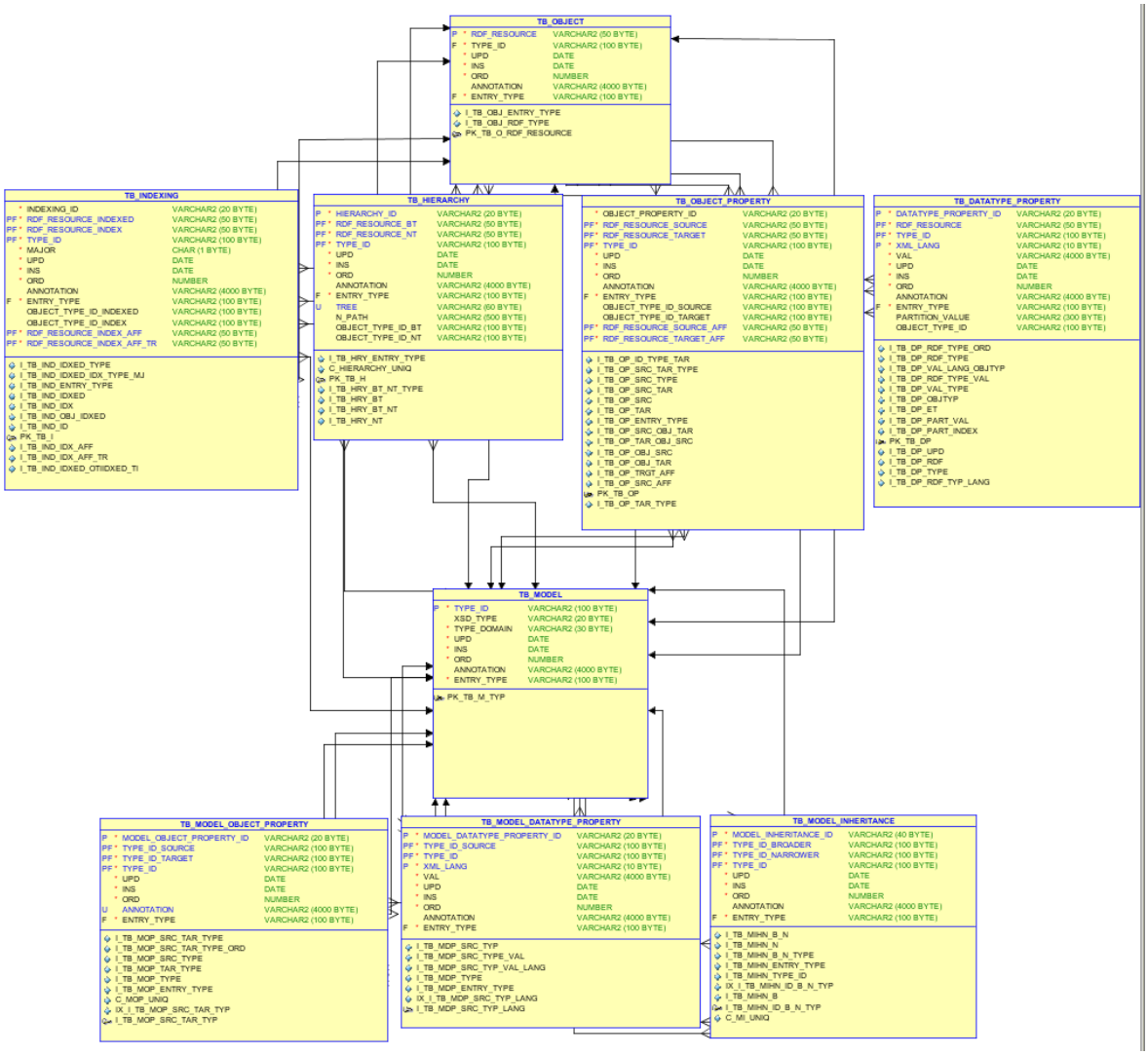

FIGURE A.1 – Le modèle physique du système d'information CISMeF

A.2 Tableaux annexes

TABLEAU A.1 – Les principales terminologies et ontologies ($n = 32$) disponibles dans le portail HeTOP (nombre de termes en français total $n = 653{,}392$)

Terminologie	Version	Termes en français	% des termes en français
Association pour le Développement de l'Informatique en Cytologie et en Anatomie Pathologiques (ADICAP)	NA	9,189	1,34 %
ATC	2014	6,005	0,88 %
Base Nationale des Cas d'Intoxications (BNCI)	2011	786	0,11 %
CCAM	v43	10,138	1,48 %
Classification Internationale du Fonctionnement, du handicap et de la santé (CIF)	2001	1,496	0,22 %
International Classification of Diseases for Oncology (ICD-O)		1,595	0,23 %
CISMeF	2016	129	0,02 %
ICPC-2	5.0	745	0,11 %
CLAssification des DIspositifs MÉDicaux (CLADIMED)	v10	5,136	0,75 %
Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV)	IV	590	0,09 %
FMA	2009	16,429	2,39 %
Genes & Proteins	2017	612	0,09 %
GO	2011	600	0,09 %
HPO	2014	11,136	1,62 %

HRDO	2013	10,689	1,56 %
CIM10	10	19,615	2,86 %
ICNP	2011	2,814	0,41 %
LOINC	2.38	58,587	8,54 %
Liste des Produits et des Prestations (LPP)	236A	5,072	0,74 %
MedDRA	v18.1	68,139	9,93 %
Medline Plus	2009	848	0,12 %
MeSH	2016	94,106	13,71 %
NCIt	2012	60,924	8,88 %
OMIM	2015	7,276	1,06 %
PASCAL	2015	6,022	0,88 %
PHARMA	2016	35,327	5,15 %
Q-Codes	2.4	188	0,03 %
SNOMED CT	2010	137,878	20,09 %
SNOMED Int.	3.5	106,266	15,49 %
SYNODOS	2015	382	0,06 %
TSP	4	7,145	1,04 %
The Unified Code for Units of Measure (UCUM)	1.9	346	0,05 %

Bibliographie

Abacha, A. B. et P. Zweigenbaum. 2011, «Medical entity recognition : a comparison of semantic and statistical methods», dans *Proceedings of BioNLP Workshop*, Association for Computational Linguistics, p. 56–64. 41

Abiteboul, S., P. Buneman et D. Suciu. 2000, *Data on the Web : from relations to semistructured data and XML*, Morgan Kaufmann. 65

Achard, F., G. Vaysseix et E. Barillot. 2001a, «XML, bioinformatics and data integration.», *Bioinformatics*, vol. 17, n o 2, p. 115–125. 65

Achard, F., G. Vaysseix et E. Barillot. 2001b, «XML, bioinformatics and data integration.», *Bioinformatics*, vol. 17, n o 2, p. 115–125. 66

Aggarwal, N., K. Asooja et P. Buitelaar. 2012, «DERI&UPM : pushing corpus based relatedness to similarity : shared task system description», dans *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume Proceedings of the main conference and the shared task, and Volume Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, p. 643–647. 137

Archer, N., U. Fevrier-Thomas, C. Lokker, K. A. McKibbon et S. E. Straus. 2011, «Personal health records : a scoping review», *Journal of the American Medical Informatics Association*, vol. 18, n o 4, p. 515–522. xv, 28, 29, 30

Aronson, A. R. et F.-M. Lang. 2010, «An overview of MetaMap : historical perspective and recent advances», *Journal of the American Medical Informatics Association*, vol. 17, n o 3, p. 229–236. 42, 44

Aronson, S. J. et H. L. Rehm. 2015, «Building the foundation for genomics in precision medicine.», *Nature*, vol. 526, n o 7573, p. 336–342. 32

Ball, M. J., N. C. Smith et R. S. Bakalar. 2007, «Personal health records : empowering consumers», *J Healthc Inf Manag*, vol. 21, p. 77. 26

Bär, D., C. Biemann, I. Gurevych et T. Zesch. 2012, «UKP : computing semantic textual similarity by combining multiple content similarity measures», dans *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume Proceedings of the main conference and the shared task, and Volume Proceedings of the Sixth International Workshop on Semantic Evaluation*, Association for Computational Linguistics, p. 435–440. 138

Barrón-Cedeño, A., P. Rosso, E. Agirre et G. Labaka. 2010, «Plagiarism detection across distant language pairs», dans *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, p. 37–45. 134

Belkin, N. J. et W. B. Croft. 1992, «Information filtering and information retrieval : two sides of the same coin ?», *Communications of the ACM*, vol. 35, n ° 12, p. 29–38. 49

Bender, O., F. J. Och et H. Ney. 2003, «Maximum entropy models for named entity recognition», dans *the seventh conference*, Association for Computational Linguistics, Morristown, NJ, USA, p. 148–151. 42

Bernstam, E. V., J. W. Smith et T. R. Johnson. 2010, «What is biomedical informatics ?», *Journal of biomedical informatics*, vol. 43, n ° 1, p. 104–110. 4

Bikel, D. M., R. Schwartz et R. M. Weischedel. 1999, «An Algorithm that Learns What's in a Name», *Machine learning*, vol. 34, n ° 1-3, p. 211–231. 42

Bird, S., E. Klein et E. Loper. 2009, *Natural Language Processing with Python*, O'Reilly Media, Inc. 145

Bodenreider, O. 2008, «Biomedical ontologies in action : role in knowledge management, data integration and decision support.», *Yearbook of medical informatics*, p. 67–79. 65

Bodenreider, O. et R. Stevens. 2006, «Bio-ontologies : current trends and future directions», *Briefings in bioinformatics*, vol. 7, n ° 3, p. 256–274. 104

Bodnari, A., L. Deleger et T. Lavergne. 2013, «A Supervised Named-Entity Extraction System for Medical Text.», *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2014)*, vol. 1179. 40

Boogerd, E. A., T. Arts, L. J. Engelen et T. H. van de Belt. 2015, «"What Is eHealth" : Time for An Update ?», *JMIR Research Protocols*, vol. 4, n ° 1, p. e29. 12

Bosworth, A. 2007, «Putting Health into the patient's Hands-Consumerism and Health care», dans *The American Medical Informatics Association AMIA Spring Congress Informatics Across the Spectrum*, Opening Plenary Session and Keynote Address, Orlando, Florida. 12

Brown, S. H., C. S. Husser, D. Wahner-Roedler, S. Bailey, L. Nugent, K. Porter, B. A. Bauer et P. L. Elkin. 2007, «Using SNOMED CT as a reference terminology to cross map two highly pre-coordinated classification systems.», *Studies in health technology and informatics*, vol. 129, n ^{o} Pt 1, p. 636–639. 104

de Bruijn, B., C. Cherry, S. Kiritchenko, J. Martin et X. Zhu. 2011, «Machine-learned solutions for three stages of clinical information extraction : the state of the art at i2b2 2010.», *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, n ^{o} 5, p. 557–562. 43

Buckley, C. et E. M. Voorhees. 2004, *Retrieval evaluation with incomplete information*, ACM, New York, New York, USA. 59

Burgun, A., P. Denier, O. Bodenreider, G. Botti, D. Delamarre et al.. 1997, «A Web terminology server using UMLS for the description of medical procedures.», *Journal of the American Medical Informatics Association*, vol. 4, n ^{o} 5, p. 356–363. 104

Buscaldi, D., R. Tournier, N. Aussenac-Gilles et J. Mothe. 2012, «IRIT : textual similarity combining conceptual similarity with an n-gram comparison method», , p. 552–556. 138

Bush, V. 1945, «As We May Think», *The Atlantic Monthly*. 48

Cabot, C., J. Grosjean, R. Lelong, A. Lefebvre, T. Lecroq, L. F. Soualmia et S. J. Darmoni. 2014, «Omic Data Modelling for Information Retrieval», dans *Proceedings of the 2nd International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO*, p. 415–424. 74

Cabot, C., R. Lelong, J. Grosjean, L. F. Soualmia et S. J. Darmoni. 2016a, «Retrieving Clinical and Omic Data from Electronic Health Records», *Stud Health Technol Inform*, vol. 221, p. 115. 107

Cabot, C., L. F. Soualmia, B. Dahamna et S. J. Darmoni. 2016b, «SIBM at CLEF eHealth Evaluation Lab 2016 : Extracting Concepts in French Medical Texts with ECMT and CIMIND», dans *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum CLEF*, p. 47–60. 105, 112, 147, 151

Cabot, C., L. F. Soualmia et S. J. Darmoni. 2016c, «Recherche d'information dans les données cliniques et omiques au sein du Dossier Patient Informatisé», dans *Journée Plateforme dIndexation* . 101

Cabot, C., L. F. Soualmia et S. J. Darmoni. 2017, «SIBM at CLEF eHealth Evaluation Lab 2017 : Multilingual Information Extraction with CIM-IND», dans *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2017)*. 147

Cabot, C., L. F. Soualmia, J. Grosjean, R. Lelong et S. J. Darmoni. 2015, «Integrating and retrieving clinical and omic data in electronic health records», cahier de recherche. 101

Cafarella, M. J., A. Halevy et J. Madhavan. 2011, «Structured data on the web», *Communications of the ACM*, vol. 54, n ^{o} 2, p. 72–79. 65

California HealthCare Foundation. 2010, «Consumers and health information technology : a national survey», URL `http://www.chcf.org/publications/2010/04/consumers-and-health-information-technology-a-national-survey`. 28

Callan, J., J. Allan, C. L. A. Clarke, S. Dumais, D. A. Evans, M. Sanderson et C. Zhai. 2007, «Meeting of the MINDS : an information retrieval research agenda», *ACM SIGIR Forum*, vol. 41, n ^{o} 2, p. 25–34. 48

Callan, J. P., W. B. Croft et J. Broglio. 1995, «TREC and TIPSTER experiments with inquery», *Information Processing & Management*, vol. 31, n ^{o} 3, p. 327–343. 55

Callan, J. P., W. B. Croft et S. M. Harding. 1992, «The INQUERY Retrieval System», dans *Proceedings of the 3rd international conference on database and expert systems applications*, édité par Springer, Springer, Vienna, Vienna, p. 78–83. 55

Cerami, E., J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer et al.. 2012, «The cBio Cancer Genomics Portal : An Open Platform for Exploring Multidimensional Cancer Genomics Data», *Cancer Discovery*, vol. 2, n ^{o} 5, p. 401–404. 34

Chapman, W. W., W. Bridewell, P. Hanbury, G. F. Cooper et B. G. Buchanan. 2001, «A simple algorithm for identifying negated findings and diseases in discharge summaries.», *Journal of biomedical informatics*, vol. 34, n ^{o} 5, p. 301–310. 43

Chebil, W., L. F. Soualmia, M. N. Omri et S. J. Darmoni. 2015, «Biomedical Concepts Extraction based on Possibilistic Network and Vector Space Model», dans *Conference on Artificial Intelligence in Medicine in Europe*, Springer, p. 227–231. 45

Chebil, W., L. F. Soualmia, M. N. Omri et S. J. Darmoni. 2016, «Indexing biomedical documents with a possibilistic network», *JASIST*, vol. 67, n º 4, p. 928–941. 45

Chen, H., S. S. Fuller, C. Friedman et W. Hersh. 2006, *Medical Informatics : Knowledge Management and Data Mining in Biomedicine*, Springer Science & Business Media. 22

Cilibrasi, R. L. et P. M. B. Vitanyi. 2007, «The Google Similarity Distance», *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, n º 3, p. 370–383. 137

Clark, C., J. Aberdeen, M. Coarr, D. Tresner-Kirsch, B. Wellner, A. Yeh et L. Hirschman. 2011, «MITRE system for clinical assertion status classification.», *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, n º 5, p. 563–567. 43

Cleverdon, C. 1967, «The Cranfield Tests on Index Language Devices», *Aslib proceedings*, vol. 19, n º 6, p. 173–194. 48

Cleverdon, C. W. 1991, *The significance of the Cranfield tests on index languages*, ACM, New York, New York, USA. 58, 59

Cogley, J., N. Stokes et J. Carthy. 2013, «Medical Disorder Recognition with Structural Support Vector Machines.», *CLEF (Working Notes)*. 41

Coletti, M. H. et H. L. Bleich. 2001, «Medical Subject Headings Used to Search the Biomedical Literature», *Journal of the American Medical Informatics Association*, vol. 8, n º 4, p. 317–323. 18

Collier, N., C. Nobata et J.-i. Tsujii. 2000, *Extracting the names of genes and gene products with a hidden Markov model*, vol. 1, Association for Computational Linguistics, Morristown, NJ, USA. 40

Collins, M. et N. Duffy. 2001, «New ranking algorithms for parsing and tagging», dans *the 40th Annual Meeting*, Association for Computational Linguistics, Morristown, NJ, USA, p. 263–270. 43

Côté, R., F. Reisinger, L. Martens, H. Barsnes, J. A. Vizcaino et H. Hermjakob. 2010, «The Ontology Lookup Service : bigger and better.», *Nucleic Acids Research*, vol. 38, n º Web Server issue, p. W155–60. 105

Croft, W. B. et D. J. Harper. 1979, «Using Probabilistic Models of Document Retrieval Without Relevance Information», *Journal of documentation*, vol. 35, n º 4, p. 285–295. 48

Dai, M. 2008, «An Efficient Solution for Mapping Free Text to Ontology Terms», dans *AMIA Summit on Translational Bioinformatics*, San Francisco, CA. 44

Darmoni, S. J., S. Pereira, A. Névéol, P. Massari, B. Dahamna, C. Letord, G. Kerdelhué, J. Piot, A. Derville et B. Thirion. 2008, «French Infobutton : an academic and business perspective.», *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, p. 920–920. 105

Darmoni, S. J., B. Thirion, J. P. Leroy, M. Douyère, B. Lacoste et al.. 2001, «Doc'CISMEF : a search tool based on "encapsulated" MeSH thesaurus», *Studies in health technology and informatics*, vol. 84, n o Pt 1, p. 314–318. 4, 15, 106

De Moor, G., M. Sundgren, D. Kalra, A. Schmidt, M. Dugas et al.. 2015, «Using electronic health records for clinical research : the case of the EHR4CR project.», *Journal of biomedical informatics*, vol. 53, p. 162–173. 33

Dehghan, A. 2013, *Boundary adjustment of events in clinical named entity recognition*, CoRR. 41

Deléger, L. et C. Grouin. 2012, *Detecting negation of medical problems in French clinical notes*, ACM, New York, New York, USA. 43

Derczynski, L., D. Maynard, G. Rizzo, M. van Erp, G. Gorrell, R. Troncy, J. Petrak et K. Bontcheva. 2015, «Analysis of named entity recognition and linking for tweets», *Information Processing & Management*, vol. 51, n o 2, p. 32–49. 132

Detmer, D., M. Bloomrosen, B. Raymond et P. Tang. 2008, «Integrated Personal Health Records : Transformative Tools for Consumer-Centric Care», *BMC Medical Informatics and Decision Making*, vol. 8, n o 1, p. 45. 29

Dexter, P. R., S. Perkins, J. M. Overhage, K. Maharry, R. B. Kohler et C. J. McDonald. 2009, «A Computerized Reminder System to Increase the Use of Preventive Care for Hospitalized Patients», *N Engl J Med*, vol. 345, n o 13, p. 965–970. 29

Dhombres, F., R. Winnenburg, J. T. Case et O. Bodenreider. 2015, «Extending the coverage of phenotypes in SNOMED CT through post-coordination.», *Studies in health technology and informatics*, vol. 216, p. 795–799. 25

Diaz, F. et D. Metzler. 2006, *Improving the estimation of relevance models using large external corpora*, ACM, New York, New York, USA. 47

Dice, L. R. 1945, «Measures of the Amount of Ecologic Association Between Species», *Ecology*, vol. 26, n o 3, p. 297–302. 135

Doan, S. et H. Xu. 2010, «Recognizing Medication related Entities in Hospital Discharge Summaries using Support Vector Machine.», *Proceedings of COLING. International Conference on Computational Linguistics*, vol. 2010, p. 259–266. 41

Douyère, M., L. F. Soualmia, A. Neveol, A. Rogozan, B. Dahamna, J.-P. Leroy, B. Thirion et S. J. Darmoni. 2004, «Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway», *Health Information and Libraries Journal*, vol. 21, n o 4, p. 253–261. 123

Dubois, D. et H. Prade. 1998, «Possibility theory : qualitative and quantitative aspects», dans *Quantified representation of uncertainty and imprecision*, Springer, p. 169–226. 56

Dupuch, M., F. Segond, A. Bittar, L. Dini et L. F. Soualmia. 2013, «Separate the grain from the chaff : make the best use of language and knowledge technologies to model textual medical data extracted from electronic health records», dans *Proceedings of the th Language Technology Conference*. 107

Egozi, O., S. Markovitch et E. Gabrilovich. 2011, «Concept-Based Information Retrieval Using Explicit Semantic Analysis», *ACM Transactions on Information Systems (TOIS)*, vol. 29, n o 2, p. 8–34. 56

Ephraim, Y. et L. R. Rabiner. 1990, «On the relations between modeling approaches for speech recognition», *IEEE Transactions on Information Theory*, vol. 36, p. 372–380. 55

Fagan, L. M. 2003, «Medical Informatics : Computer Applications in Health Care and Biomedicine (Health Informatics)», *(30 April 2003)*. 13, 15, 17

Fan, J., N. Sood et Y. Huang. 2013, «Disorder concept identification from clinical notes an experience with the ShARe/CLEF 2013 challenge», . 42

Fernald, G. H., E. Capriotti, R. Daneshjou, K. J. Karczewski et R. B. Altman. 2011, «Bioinformatics challenges for personalized medicine», *Bioinformatics*, vol. 27, n o 13, p. 1741–1748. 31

Finkel, J., S. Dingare, C. D. Manning, M. Nissim, B. Alex et C. Grover. 2005, «Exploring the boundaries : gene and protein identification in biomedical text.», *BMC bioinformatics*, vol. 6 Suppl 1, n o Suppl 1, p. S5. 40

Finkel, J., S. Dingare, H. Nguyen, M. Nissim, C. Manning et G. Sinclair. 2004, «Exploiting context for biomedical entity recognition : from syntax to the web», , p. 88–91. 40

Fox, E. A., S. Betrabet, M. Koushik et W. C. Lee. 1992, «Extended Boolean Models.», . 50

Fox, E. A. et S. Sharan. 1986, « A Comparison of Two Methods For Soft Boolean Operator Interpretation In Information Retrieval», . 50

Frakes, W. B. et R. Baeza-Yates. 1992, «Information Retrieval : Data Structures and Algorithms», *(12 June 1992)*. 49

Frixione, M. et A. Lieto. 2012, «Representing concepts in formal ontologies. Compositionality vs. typicality effects», *Logic and Logical Philosophy*, vol. 21, n ^{o} 4, p. 391–414. 24

Gabrilovich, E. et S. Markovitch. 2007, «Computing semantic relatedness using wikipedia-based explicit semantic analysis.», *IJcAI*. 136

Gaines, B. R. et T. L. Kohout. 1975, «Possible automata», dans *Proc. Int. Symp. Multiple-Valued logics*, Citeseer, Bloomington, IN, p. 183–196. 55

Gambarte, M. L., A. L. Osornio, M. Martinez, G. Reynoso, D. Luna et F. G. B. de Quiros. 2007, «A practical approach to advanced terminology services in health information systems.», *Studies in health technology and informatics*, vol. 129, n ^{o} Pt 1, p. 621–625. 104

Ganslandt, T., S. Mate, K. Helbing, U. Sax et H. U. Prokosch. 2011, «Unlocking Data for Clinical Research–The German i2b2 Experience», *Applied clinical informatics*, vol. 2, n ^{o} 1, p. 116. 93

Garcelon, N., R. Salomon et A. Burgun. 2014, «Enrichissement sémantique associé à la détection de la négation et des antécédents familiaux dans un entrepôt de données hospitalier.», *JFIM*. 43

Gindl, S., K. Kaiser et S. Miksch. 2008, «Syntactical negation detection in clinical practice guidelines.», *Studies in health technology and informatics*, vol. 136, p. 187–192. 43

Goeuriot, L., L. Kelly, H. Suominen, L. Hanlen, A. Neveol, C. Grouin, J. Palotti et G. Zuccon. 2015, «Overview of the CLEF eHealth Evaluation Lab 2015», dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Cham, Cham, p. 429–443. 107

Goeuriot, L., L. Kelly, H. Suominen, A. Neveol, A. Robert, E. Kanoulas, R. Spijker, J. Palotti et G. Zuccon. 2017, «CLEF 2017 eHealth Evaluation Lab Overview», dans *CLEF - 8th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science LNCS, Springer*. 147

Griffon, N. 2013, *Modélisation, création et évaluation de flux de terminologies et de terminologies d'interface : application à la production d'examens complémentaires de biologie et d'imagerie médicale.*, thèse de doctorat, Université de Rouen. 27

Griffon, N., M. Schuers et S. J. Darmoni. 2016, «[LiSSa : An alternative in French to browse health scientific literature ?]», *Presse medicale (Paris, France : 1983)*, vol. 45, n o 11, p. 955–956. 15

Griffon, N., M. Schuers, L. F. Soualmia, J. Grosjean, G. Kerdelhué, I. Kergourlay, B. Dahamna et S. J. Darmoni. 2014, «A Search Engine to Access PubMed Monolingual Subsets : Proof of Concept and Evaluation in French», *Journal of Medical Internet Research*, vol. 16, n o 12, p. e271. 120

Grosjean, J. 2014, «Modélisation, réalisation et évaluation d'un portail multi-terminologique multi-discipline, multi-lingue (3M) dans le cadre de la Plateforme d'Indexation Régionale (PlaIR)», *www.theses.fr*. 2, 80, 105

Grosjean, J., T. Merabti, B. Dahamna, I. Kergourlay, B. Thirion, L. F. Soualmia et S. J. Darmoni. 2011, «Health multi-terminology portal : a semantic added-value for patient safety.», *Studies in health technology and informatics*, vol. 166, p. 129–138. 4, 105

Grouin, C. 2014, «Biomedical entity extraction using machine-learning based approaches», *substance*. 40

Gruber, T. R. 1993, «A translation approach to portable ontology specifications», *Knowledge acquisition*. 18

Gruber, T. R. 1995, «Toward principles for the design of ontologies used for knowledge sharing ?», *International journal of human-computer studies*, vol. 43, n o 5-6, p. 907–928. 18

Halamka, J. D., K. D. Mandl et P. C. Tang. 2008, «Early Experiences with Personal Health Records», *Journal of the American Medical Informatics Association*, vol. 15, n o 1, p. 1–7. 28

Hall, P. A. V. et G. R. Dowling. 1980, «Approximate String Matching», *ACM computing surveys (CSUR)*, vol. 12, n o 4, p. 381–402. 134

Hamid, J. S., P. Hu, N. M. Roslin, V. Ling, C. M. T. Greenwood et J. Beyene. 2009, «Data integration in genetics and genomics : methods and challenges.», *Human genomics and proteomics : HGP*, vol. 2009, n o 2, p. 1–13. 64

Hamosh, A., A. F. Scott, J. S. Amberger, C. A. Bocchini et V. A. McKusick. 2005, «Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.», *Nucleic Acids Research*, vol. 33, n ° Database issue, p. D514–7. 69

Harkema, H., J. N. Dowling, T. Thornblade et W. W. Chapman. 2009, «ConText : an algorithm for determining negation, experiencer, and temporal status from clinical reports.», *Journal of biomedical informatics*, vol. 42, n ° 5, p. 839–851. 43

Hendler, J. 2014, «Data Integration for Heterogenous Datasets.», *Big data*, vol. 2, n ° 4, p. 205–215. 65

Hersh, W. 2008, *Information Retrieval : A Health and Biomedical Perspective*, Health Informatics, Springer Science & Business Media, New York, NY. 12, 13

Hettne, K. M., E. M. van Mulligen, M. J. Schuemie, B. J. Schijvenaars et J. A. Kors. 2010, «Rewriting and suppressing UMLS terms for improved biomedical term identification», *Journal of biomedical semantics*, vol. 1, n ° 1, p. 5. 45

Hiemstra, D. et W. Kraaij. 1998, *Twenty-One in ad-hoc and CLIR*, Proc. of TREC-7. 55

Hoerbst, A. et E. Ammenwerth. 2010, «Electronic health records», *Methods of Information in Medicine*, vol. 49, p. 320–336. 27

Hood, D. 2004, «Caverphone revisited», *Technical Paper CTP150804*. 140

Humphreys, B. L., D. A. Lindberg, H. M. Schoolman et G. O. Barnett. 1998, «The Unified Medical Language System : an informatics research collaboration.», *Journal of the American Medical Informatics Association*, vol. 5, n ° 1, p. 1–11. 20

Iakovidis, I. 1998, «Towards personal health record : current situation, obstacles and trends in implementation of electronic healthcare record in Europe1», *International Journal of Medical Informatics*, vol. 52, n ° 1, p. 105–115. 27

Islam, A. et D. Inkpen. 2008, «Semantic text similarity using corpus-based word similarity and string similarity», *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 2, n ° 2, p. 10–25. 137

Iyer, S. V., R. Harpaz, P. LePendu, A. Bauer-Mehren et N. H. Shah. 2014, «Mining clinical text for signals of adverse drug-drug interactions.», *Journal of the American Medical Informatics Association : JAMIA*, vol. 21, n ° 2, p. 353–362. 33

Jaccard, P. 1901, «Etude comparative de la distribution florale dans une portion des Alpes et des Jura», *Bull Soc Vaudoise Sci Nat*, vol. 37, p. 547–579. 135

Jamoulle, M. 2013, «Using the International Classification for Primary Care (ICPC) and the Core Content Classification for General Practice (3CGP) to classify conference abstracts», *Rev Port Med Geral Fam*, vol. 29, n º 5, p. 66–67. 123

Jaro, M. A. 1995, «Probabilistic linkage of large public health data files», *Statistics in medicine*, vol. 14, n º 5-7, p. 491–498. 134

Jaro, M. A. 2012, «Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida», *Journal of the American Statistical Association*, vol. 84, n º 406, p. 414–420. 134

Jensen, F. V. et F. B. Jensen. 2001, *Bayesian Networks and Decision Graphs*, Springer. 54

Jimeno, A., E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga et D. Rebholz-Schuhmann. 2008, «Assessment of disease named entity recognition on a corpus of annotated sentences.», *BMC bioinformatics*, vol. 9 Suppl 3, n º Suppl 3, p. S3. 41

Joachims, T., L. Granka, B. Pan, H. Hembrooke et G. Gay. 2005, *Accurately interpreting clickthrough data as implicit feedback*, ACM, New York, New York, USA. 53

Johnson, E. K., S. Broder-Fingert, P. Tanpowpong, J. Bickel, J. R. Lightdale et C. P. Nelson. 2014, «Use of the i2b2 research query tool to conduct a matched case–control clinical research study : advantages, disadvantages and methodological considerations», *BMC medical research methodology*, vol. 14, n º 1, p. 16. 33

Jones, K. S. 1972, «A Statistical Interpretation of Term Specificity and Its Application in Retrieval», *Journal of documentation*, vol. 28, n º 1, p. 11–21. 48

Jonquet, C., N. Shah et M. Musen. 2009, «The Open Biomedical Annotator», dans *Summit on translational bioinformatics*, p. 56–60. 44

Jydstrup, R. A. et M. J. Gross. 1966, «Cost of information handling in hospitals.», *Health services research*, vol. 1, n º 3, p. 235–271. 12

Karlsson, D., M. Nyström et R. Cornet. 2014, «Does SNOMED CT post-coordination scale ?», *Studies in health technology and informatics*, vol. 205, p. 1048–1052. 25

Kelly, L., L. Goeuriot, H. Suominen, A. Neveol, J. Palotti et G. Zuccon. 2016, «Overview of the CLEF eHealth Evaluation Lab 2016», dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Springer, Cham, p. 255–266. 61, 107, 147

Kleinsorge, R. et J. Willis. 2008, «Unified Medical Language System (UMLS) Basics», *National Library of Medicine*. xiii, 20, 21, 22

Kohane, I. S. 2011, «Using electronic health records to drive discovery in disease genomics», *Nature Reviews Genetics*, vol. 12, n o 6, p. 417–428. 33

Komatsoulis, G. A., D. B. Warzel, F. W. Hartel, K. Shanbhag, R. Chilukuri et al.. 2008, «caCORE version 3 : Implementation of a model driven, service-oriented architecture for semantic interoperability.», *Journal of biomedical informatics*, vol. 41, n o 1, p. 106–123. 104

Koopman, B., P. D. Bruza, L. Sitbon et M. Lawley. 2011, «Towards semantic search and inference in electronic medical records : an approach using concept-based information retrieval», *School of Information Systems ; Science & Engineering Faculty*. 56

Korkontzelos, I., D. Piliouras, A. W. Dowsey et S. Ananiadou. 2015, «Boosting drug named entity recognition using an aggregate classifier.», *Artificial Intelligence in Medicine*, vol. 65, n o 2, p. 145–153. 41

Krause, E. F. 1973, «Taxicab geometry», *The Mathematics Teacher*. 134

Lafferty, J., A. McCallum et F. Pereira. 2001, «Conditional random fields : Probabilistic models for segmenting and labeling sequence data», dans *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning*, p. 282–289. 40

Lai, K. H., M. Topaz, F. R. Goss et L. Zhou. 2015, «Automated misspelling detection and correction in clinical free-text records.», *Journal of biomedical informatics*, vol. 55, p. 188–195. 132

Lancaster, F. W. W. A. J. 1993, *Information Retrieval Today. Revised, Retitled, and Expanded Edition.*, Information Resources Press. 49

Landauer, T. K. et S. T. Dumais. 1997, «A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge.», *Psychological review*, vol. 104, n o 2, p. 211–240. 136

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody et al.. 2001, «Initial sequencing and analysis of the human genome.», *Nature*, vol. 409, n° 6822, p. 860–921. 16

Lelong, R., C. Cabot et L. F. Soualmia. 2016, «Semantic Search Engine to Query into Electronic Health Records with a Multiple-Layer Query Language», dans *Proceedings of the 2nd SIGIR workshop on Medical Information Retrieval (MedIR)*. 88, 107

Levenshtein, V. I. 1966, «Binary codes capable of correcting deletions, insertions, and reversals», dans *Soviet physics doklady*, p. 707–710. 134

Li, Y., D. McLean, Z. A. Bandar, J. D. O'shea et K. Crockett. 2006, «Sentence similarity based on semantic nets and corpus statistics», *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, n° 8, p. 1138–1150. 137

Lin, J. et W. J. Wilbur. 2007, «PubMed related articles : a probabilistic topic-based model for content similarity», *BMC bioinformatics*, vol. 8, n° 1, p. 423. 44

Lindberg, D. A., B. L. Humphreys et A. T. McCray. 1993, «The Unified Medical Language System.», *Methods of Information in Medicine*, vol. 32, n° 4, p. 281–291. 20

Liu, Z. et W. W. Chu. 2007, «Knowledge-based query expansion to support scenario-specific retrieval of medical free text», *Information Retrieval*, vol. 10, n° 2, p. 173–202. 56, 57

Lowe, H. J., T. A. Ferris, P. M. Hernandez et S. C. Weber. 2009, «STRIDE–An integrated standards-based translational research informatics platform.», *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2009, p. 391–395. 33

Luhn, H. P. 1957, «A statistical approach to mechanized encoding and searching of literary information», *IBM Journal of Research and Development*. 48

Lund, K. et C. Burgess. 1996, «Producing high-dimensional semantic spaces from lexical co-occurrence», *Behavior Research Methods, Instruments, & Computers*, vol. 28, n° 2, p. 203–208. 135

Lund, K., C. Burgess et R. A. Atchley. 1995, «Semantic and associative priming in high-dimensional semantic space», dans *Lund, K., Burgess, C., Atchley, R. A. , July. Semantic and associative priming in high-dimensional semantic space. In Proceedings of the th annual conference of the Cognitive Science Society*, Proceedings of the 17th annual conference of the ..., p. 660–665. 135

Lyman, P. et H. R. Varian. 2003, «How Much Information», URL http://groups.ischool.berkeley.edu/archive/how-much-info-2003/. 12

Ma, Y., J.-j. Kim, B. Bigot et T. M. Khan. 2016, «Feature-enriched word embeddings for named entity recognition in open-domain conversations», dans *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 6055–6059. 132

Madhavan, S., Y. Gusev, M. Harris, D. M. Tanenbaum, R. Gauba et al.. 2011, «G-DOC : A Systems Medicine Platform for Personalized Oncology», *Neoplasia*, vol. 13, n o 9, p. 771–783. 35

Maglott, D., J. Ostell, K. D. Pruitt et T. Tatusova. 2007, «Entrez Gene : gene-centered information at NCBI», *Nucleic Acids Research*, vol. 35, p. D26–D31. 68

Mamlin, J. J. et D. H. Baker. 1973, «Combined time-motion and work sampling study in a general medicine clinic.», *Medical care*, vol. 11, n o 5, p. 449–456. 12

Manning, C. D., P. Raghavan et H. Schütze. 2008, «Introduction to information retrieval», . 58

Marcus, R. S. 1991, «Computer and Human Understanding in Intelligent Retrieval Assistance.», *Proceedings of the ASIS Annual Meeting*, vol. 28, p. 49–59. 49

Maron, M. E. et J. L. Kuhns. 1960, «On Relevance, Probabilistic Indexing and Information Retrieval», *Journal of the ACM (JACM)*, vol. 7, n o 3, p. 216–244. 53

Mary, M. 2017, *Interopérabilité sémantique pour les données de diagnostic in vitro : représentation des connaissances et alignement*, thèse de doctorat, Université de Rouen Normandie. 27

Matveeva, I., G. A. Levow, A. Farahat et C. Royer. 2005, *Generalized latent semantic analysis for term representation*, Proc. of RANLP. 136

McCallum, A. et W. Li. 2003, «Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons», dans *the seventh conference*, Association for Computational Linguistics, Morristown, NJ, USA, p. 188–191. 42

McConnell, P., R. C. Dash, R. Chilukuri, R. Pietrobon, K. Johnson, R. Annechiarico et A. J. Cuticchia. 2008, «The cancer translational research informatics platform», *BMC Medical Informatics and Decision Making*, vol. 8, n o 1, p. 60. 34

McDonald, R. et F. Pereira. 2005, «Identifying gene and protein mentions in text using conditional random fields.», *BMC bioinformatics*, vol. 6 Suppl 1, n ° Suppl 1, p. S6. 40

Meij, E., D. Trieschnigg, M. de Rijke et W. Kraaij. 2010, «Conceptual language models for domain-specific retrieval», *Information Processing & Management*, vol. 46, n ° 4, p. 448–469. 57

Menasalvas, E. et C. Gonzalo-Martin. 2016, «Challenges of Medical Text and Image Processing : Machine Learning Approaches», dans *Machine Learning for Health Informatics*, Springer International Publishing, Cham, p. 221–242. 132

Miyoshi, N. S. B., D. G. Pinheiro, W. A. Silva Jr et J. C. Felipe. 2013, «Computational framework to support integration of biomolecular and clinical data within a translational approach», *BMC bioinformatics*, vol. 14, p. 180. 33

Mooers, C. S. 1950, «Coding, Information Retrieval, and the Rapid Selector», *American Documentation*. 48

Mork, J., A. Aronson et D. Demner Fushman. 2017, «12 years on - Is the NLM medical text indexer still useful and relevant ?», *Journal of biomedical semantics*, vol. 8, n ° 1, p. 8. 42, 44, 132

Mork, J. G., D. Demner-Fushman et S. Schmidt. 2014, «Recent Enhancements to the NLM Medical Text Indexer.», *CEUR-WS Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2014)*, p. 1328–1336. 44

Moschitti, A. 2004, «A study on convolution kernels for shallow semantic parsing», dans *the 42nd Annual Meeting*, Association for Computational Linguistics, Morristown, NJ, USA, p. 335–es. 43

Murphy, S. N., P. Avillach, R. Bellazzi, L. Phillips, M. Gabetta, A. Eran, M. T. McDuffie et I. S. Kohane. 2017, «Combining clinical and genomics queries using i2b2 - Three methods.», *PloS one*, vol. 12, n ° 4, p. e0172 187. 33, 94

Murphy, S. N., M. E. Mendis, D. A. Berkowitz, I. Kohane et H. C. Chueh. 2006, «Integration of clinical and genetic data in the i2b2 architecture.», *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2006, p. 1040. 33, 93, 94

Musen, M. A., N. F. Noy, N. H. Shah, P. L. Whetzel, C. G. Chute, M.-A. Story et B. Smith. 2012, «The National Center for Biomedical Ontology», *Journal of the American Medical Informatics Association*, vol. 19, n ° 2, p. 190–195. 44

Nadeau, D. et S. Sekine. 2007, «A survey of named entity recognition and classification», *Lingvisticae Investigationes*, vol. 30, n ° 1, p. 3–26. 42

Navas, H., A. L. Osornio, A. Baum, A. Gomez, D. Luna et F. G. B. de Quiros. 2007, «Creation and evaluation of a terminology server for the interactive coding of discharge summaries.», *Studies in health technology and informatics*, vol. 129, n ° Pt 1, p. 650–654. 104

Needleman, S. B. et C. D. Wunsch. 1970, «A general method applicable to the search for similarities in the amino acid sequence of two proteins», *Journal of molecular biology*, vol. 48, n ° 3, p. 443–453. 134

Névéol, A., R. N. Anderson, K. B. Cohen et C. Grouin. 2017, «CLEF eHealth 2017 Multilingual Information Extraction task overview : ICD10 coding of death certificates in English and French», *CLEF 2017 Evaluation Forum*. xvi, 149, 150, 151, 152

Névéol, A., L. Goeuriot et L. Kelly. 2016, «Clinical information extraction at the CLEF eHealth evaluation lab 2016», dans *Proceedings of the CLEF 2015 Evaluation Labs and Workshop : Online Working Notes*. xvi, 45, 107, 147, 148

Névéol, A., J. Grosjean, S. J. Darmoni et P. Zweigenbaum. 2014, «Language Resources for French in the Biomedical Domain.», *LREC*. 120

Neveol, A., C. Grouin, J. Leixa, S. Rosset et P. Zweigenbaum. 2014a, «The Quaero French medical corpus : A ressource for medical entity recognition and normalization», . 113

Neveol, A., C. Grouin, J. Leixa, S. Rosset et P. Zweigenbaum. 2014b, «The Quaero French medical corpus : A ressource for medical entity recognition and normalization», . 147

Névéol, A., C. Grouin et X. Tannier. 2015, «CLEF eHealth Evaluation Lab 2015 Task 1b : Clinical Named Entity Recognition.», *CLEF 2015 Working Notes*. 107

Ng, S. B., K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor et al.. 2010, «Exome sequencing identifies the cause of a mendelian disorder.», *Nature genetics*, vol. 42, n ° 1, p. 30–35. 31

Ohno-Machado, L., V. Bafna, A. A. Boxwala, B. E. Chapman, W. W. Chapman et al.. 2012, «iDASH : integrating data for analysis, anonymization, and sharing», *Journal of the American Medical Informatics Association*, vol. 19, n ° 2, p. 196–201. 35

Patel, C., J. Cimino, J. Dolby, A. Fokoue, A. Kalyanpur, A. Kershenbaum, L. Ma, E. Schonberg et K. Srinivas. 2007, «Matching Patient Records to Clinical Trials Using Ontologies», dans *The Semantic Web*, Springer Berlin Heidelberg, Berlin, Heidelberg, p. 816–829. 11

Patrick, J., Y. Wang et P. Budd. 2006, «Automatic mapping clinical notes to medical terminologies», *Australasian language technology workshop*. 43

Pavillon, G. et F. Laurent. 2003, «Certification et codification des causes médicales de décès», *Bulletin épidémiologique hebdomadaire*. 141

Pereira, S., A. Névéol, G. Kerdelhué et E. Serrot. 2008, «Using multiterminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue.», *AMIA*. 45, 105

Peterson, J. L. 1980, «Computer programs for detecting and correcting spelling errors», *Communications of the ACM*, vol. 23, n ^{o} 12, p. 676–687. 134

Philips, L. 2000, «The double metaphone search algorithm», *C/C++ Users Journal*, vol. 18, n ^{o} 6, p. 38–43. 140, 145

Ponte, J. M. et W. B. Croft. 1998, *A language modeling approach to information retrieval*, ACM, New York, New York, USA. 55

Potthast, M., B. Stein et M. Anderka. 2008, «A Wikipedia-based multilingual retrieval model», *Advances in Information Retrieval*. 136

Proux, D., F. Rechenmann, L. Julliard, V. Pillet et B. Jacq. 1998, «Detecting Gene Symbols and Names in Biological Texts : A First Step toward Pertinent Information Extraction.», *Genome informatics. Workshop on Genome Informatics*, vol. 9, p. 72–80. 41

Ratinov, L. et D. Roth. 2009, *Design challenges and misconceptions in named entity recognition*, Association for Computational Linguistics. 42

Ravindran, D. et S. Gauch. 2004, *Exploiting hierarchical relationships in conceptual search*, ACM, New York, New York, USA. 56

Rebholz-Schuhmann, D., H. Kirsch, S. Gaudan, M. Arregui et G. Nenadic. 2006, «Annotation and disambiguation of semantic types in biomedical text : a cascaded approach to named entity recognition», , p. 11–18. 41

Robertson, S. E. 1977, «The Probability Ranking Principle in IR», *Journal of documentation*, vol. 33, n ^{o} 4, p. 294–304. 46, 52

Robertson, S. E. et K. S. Jones. 1976, «Relevance weighting of search terms», *Journal of the Association for Information Science and Technology*, vol. 27, n º 3, p. 129–146. 48, 53

Robertson, S. E. et S. Walker. 1994, *Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval*, Springer-Verlag New York, Inc. 53

Rubin, D. D. L., S. E. Lewis, C. J. Mungall, S. Misra, M. Westerfield et al.. 2006, «National Center for Biomedical Ontology : Advancing Biomedicine through Structured Organization of Scientific Knowledge», *www.liebertpub.com*, vol. 10, n º 2, p. 185–198. 105

Rubin, D. L., N. H. Shah et N. F. Noy. 2008, «Biomedical ontologies : a functional perspective», *Briefings in bioinformatics*, vol. 9, n º 1, p. 75–90. 104

Sachs, W. M. 1976, «An approach to associative retrieval through the theory of fuzzy sets», *Journal of the Association for Information Science and Technology*, vol. 27, n º 2, p. 85–87. 50

Saha, S. K., S. Sarkar et P. Mitra. 2009, «Feature selection techniques for maximum entropy based biomedical named entity recognition», *Journal of biomedical informatics*, vol. 42, n º 5, p. 905–911. 40, 41

Salton, G. 1968, «Automatic information organization and retrieval», . 46

Salton, G. 1971, «The SMART Retrieval System—Experiments in Automatic Document Processing», . 48, 50, 52

Salton, G., E. A. Fox et H. Wu. 1983, «Extended Boolean information retrieval», *Communications of the ACM*, vol. 26, n º 11, p. 1022–1036. 50

Salton, G. et C. S. Yang. 1973, «On The Specification Of Term Values In Automatic Indexing», *Journal of documentation*, vol. 29, n º 4, p. 351–372. 52

Sarkar, I. N., A. J. Butte, Y. A. Lussier, P. Tarczy-Hornoch et L. Ohno-Machado. 2011, «Translational bioinformatics : linking knowledge across biological and clinical realms.», *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, n º 4, p. 354–357. 32

Scheufele, E., D. Aronzon, R. Coopersmith, M. T. McDuffie, M. Kapoor, C. A. Uhrich, J. E. Avitabile, J. Liu, D. Housman et M. B. Palchuk. 2014, «tranSMART : An Open Source Knowledge Management and High Content Data Analytics Platform.», *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2014, p. 96–101. 95

Settles, B. 2004, «Biomedical named entity recognition using conditional random fields and rich feature sets», , p. 104–107. 40

Shah, N. H., N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang et M. A. Musen. 2009a, «Comparison of concept recognizers for building the Open Biomedical Annotator», *BMC bioinformatics*, vol. 10, n ° 9, p. S14. 44

Shah, N. H., C. Jonquet, A. P. Chiang, A. J. Butte, R. Chen et M. A. Musen. 2009b, «Ontology-driven indexing of public datasets for translational bioinformatics», *BMC bioinformatics*, vol. 10, n ° 2, p. S1. 44

Sherry, S. T., M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski et K. Sirotkin. 2001, «dbSNP : the NCBI database of genetic variation», *Nucleic Acids Research*, vol. 29, n ° 1, p. 308–311. 91

Shimokawa, K., K. Mogushi, S. Shoji, A. Hiraishi, K. Ido, H. Mizushima et H. Tanaka. 2010, «iCOD : an integrated clinical omics database based on the systems-pathology view of disease», *BMC Genomics*, vol. 11, n ° 4, p. S19. 35

Sidorov, G., F. Velasquez, E. Stamatatos, A. Gelbukh et L. Chanona-Hernández. 2012, «Syntactic Dependency-Based N-grams as Classification Features», dans *Advances in Computational Intelligence*, Springer, Berlin, Heidelberg, Berlin, Heidelberg, p. 1–11. 43

Siklósi, B., A. Novák et G. Prószéky. 2016, «Context-aware correction of spelling errors in Hungarian medical documents», *Computer Speech & Language*, vol. 35, p. 219–233. 132

Smith, T. F. et M. S. Waterman. 1981, «Identification of common molecular subsequences», *Journal of molecular biology*, vol. 147, n ° 1, p. 195–197. 134

Sohn, S., S. Wu et C. G. Chute. 2012, «Dependency Parser-based Negation Detection in Clinical Narratives.», *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2012, p. 1–8. 43

Soualmia, L. F., C. Cabot, B. Dahamna et S. J. Darmoni. 2015, «SIBM at CLEF e-Health Evaluation Lab 2015», dans *Proceedings of the CLEF 2015 Evaluation Labs and Workshop : Online Working Notes*. 105, 112

Soualmia, L. F. et S. J. Darmoni. 2005, «Combining different standards and different approaches for health information retrieval in a quality-controlled gateway», *International Journal of Medical Informatics*, vol. 74, n ° 2-4, p. 141–150. 104

Spackman, K. 2008, «SNOMED clinical terms basics», cahier de recherche. 24

Sparck Jones, K., S. Walker et S. E. Robertson. 2000, «A probabilistic model of information retrieval : development and comparative experiments», *Information Processing & Management*, vol. 36, n ° 6, p. 809–840. 54

Spasić, I., B. Zhao, C. B. Jones et K. Button. 2015, «KneeTex : an ontology-driven system for information extraction from MRI reports.», *Journal of biomedical semantics*, vol. 6, n ° 1, p. 34. 42

Stein, L. D. 2003, «Integrating biological databases.», *Nature Reviews Genetics*, vol. 4, n ° 5, p. 337–345. 65

Stokes, N., Y. Li, L. Cavedon et J. Zobel. 2009, «Exploring criteria for successful query expansion in the genomic domain», *Information Retrieval*, vol. 12, n ° 1, p. 17–50. 57

Stolyar, A., W. B. Lober, D. R. Drozd et J. Sibley. 2005, «Feasibility of data exchange with a Patient-centered Health Record.», *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2005, p. 1123. 28

Sun, W., A. Rumshisky et Ö. Uzuner. 2013, «Evaluating temporal relations in clinical text : 2012 i2b2 Challenge», *Journal of the American Medical Informatics Association*, vol. 20, n ° 5, p. 806–813. 62

Szalma, S., V. Koka, T. Khasanova et E. D. Perakslis. 2010, «Effective knowledge management in translational medicine», *Journal of Translational Medicine*, vol. 8, n ° 1, p. 68. 36, 95

Takai-Igarashi, T., R. Akasaka, K. Suzuki, T. Furukawa, M. Yoshida et al.. 2011, «On experiences of i2b2 (Informatics for integrating biology and the bedside) database with Japanese clinical patients' data», *Bioinformation*, vol. 6, n ° 2, p. 86–90. 93, 95

Tan, A., B. Tripp et D. Daley. 2011, «BRISK—research-oriented storage kit for biology-related data», *Bioinformatics*, vol. 27, n ° 17, p. 2422–2425. 34

Tang, B., H. Cao, X. Wang, Q. Chen et H. Xu. 2014, «Evaluating word representation features in biomedical named entity recognition tasks.», *BioMed research international*, vol. 2014, n ° 2, p. 240 403–6. 40

Tang, L., S. Rajan et V. K. Narayanan. 2009, *Large scale multi-label classification via metalabeler*, ACM, New York, New York, USA. 44

Tang, P. C., J. S. Ash, D. W. Bates, J. M. Overhage et D. Z. Sands. 2006, «Personal Health Records : Definitions, Benefits, and Strategies for Overcoming Barriers to Adoption», *Journal of the American Medical Informatics Association*, vol. 13, n° 2, p. 121–126. 29

The Gene Ontology Consortium. 2008, «The Gene Ontology project in 2008», *Nucleic Acids Research*, vol. 36, p. D440–D444. 68

Thiessard, F., F. Mougin, G. Diallo, V. Jouhet, S. Cossin et al.. 2012, «RAVEL : retrieval and visualization in ELectronic health records.», *Studies in health technology and informatics*, vol. 180, p. 194–198. 3, 107

Tkachenko, M. et A. Simanovsky. 2012, «Named entity recognition : Exploring features.», dans *KONVENS*, p. 118–127. 42

Trieschnigg, D. 2010, *Proof of concept : concept-based biomedical information retrieval*, thèse de doctorat, Centre for Telematics and Information Technology, University of Twente, Enschede, The Netherlands. 56, 57

Trieschnigg, D., D. Hiemstra, F. de Jong et W. Kraaij. 2010, «A cross-lingual framework for monolingual biomedical information retrieval», dans *Proceedings of the 19th ACM international conference*, ACM, p. 169–178. 56

Tsoumakas, G., M. Laliotis, N. Markantonatos et I. Vlahavas. 2013, «Large-scale semantic indexing of biomedical publications at bioasq», dans *BioASQ Workshop*. 44

Turchin, A., N. S. Kolatkar, R. W. Grant, E. C. Makhni, M. L. Pendergrass et J. S. Einbinder. 2006, «Using Regular Expressions to Abstract Blood Pressure and Treatment Intensification Information from the Text of Physician Notes», *Journal of the American Medical Informatics Association*, vol. 13, n° 6, p. 691–695. 66

Turtle, H. et W. B. Croft. 1989, «Inference networks for document retrieval», dans *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, New York, USA, p. 1–24. 54

Turtle, H. et W. B. Croft. 1991, «Evaluation of an inference network-based retrieval model», *ACM Transactions on Information Systems (TOIS)*, vol. 9, n° 3, p. 187–222. 48, 54

UniProt Consortium. 2013, «Update on activities at the Universal Protein Resource (UniProt) in 2013.», *Nucleic Acids Research*, vol. 41, n° Database issue, p. D43–7. 68

Uzuner, Ö., B. R. South, S. Shen et S. L. DuVall. 2011, «2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.», *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, n ° 5, p. 552–556. 43

Valentine, P. M. 2012, *A Social History of Books and Libraries from Cuneiform to Bytes*, Scarecrow Press. 47

Van Mulligen, E., Z. Afzal, S. A. Akhondi, D. Vo et J. A. Kors. 2016, «Erasmus MC at CLEF eHealth 2016 : Concept recognition and coding in French texts», . 45

Van Rijsbergen, C. J. 1986, «A non-classical logic for information retrieval», *The computer journal*, vol. 29, n ° 6, p. 481–485. 46

Vanopstal, K., R. Vander Stichele, G. Laureys et J. Buysschaert. 2011, «Vocabularies and retrieval tools in biomedicine : disentangling the terminological knot.», *Journal of medical systems*, vol. 35, n ° 4, p. 527–543. 18

Viangteeravat, T., I. M. Brooks, E. J. Smith, N. Furlotte, S. Vuthipadadon, R. Reynolds et C. S. McDonald. 2009, «Slim-prim : a biomedical informatics database to promote translational research.», *Perspectives in health information management*, vol. 6, p. 6. 33

Voorhees, E. M. 1994, «Query Expansion using Lexical-Semantic Relations», dans *SIGIR '94*, Springer London, London, p. 61–69. 56

Voorhees, E. M. et D. K. Harman. 2005, «TREC : Experiment and evaluation in information retrieval», Cambridge : MIT press. 60

Voorhees, E. M. et W. R. Hersh. 2012, «Overview of the TREC 2012 Medical Records Track.», *NIST Special Publication*. 58, 61

Waegemann, C. P. 2002, *Status report 2002 : electronic health records*, Medical Records Institute, Chicago, Illinois. 26

Waegemann, C. P. 2003, «Ehr vs. cpr vs. emr», *Healthcare Informatics online*. 26

Wang, Y. et J. Patrick. 2009, «Cascading classifiers for named entity recognition in clinical notes», dans *Proceedings of the workshop on biomedical information extraction*, Association for Computational Linguistics, p. 42–49. 41

Webster, F. 2014, *Theories of the Information Society*, Routledge. 11

Wehling, M. 2015, *Principles of Translational Science in Medicine*, Academic Press. 5

Wilbur, W. J., G. F. Hazard, G. Divita, J. G. Mork, A. R. Aronson et A. C. Browne. 1999, «Analysis of biomedical text for chemical names : a comparison of three methods.», dans *Proceedings of the AMIA Symposium*, National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894, USA., American Medical Informatics Association, p. 176–180. 41

Winkler, W. E. 1990, «String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.», dans *Proceedings of the Section on Survey Research*, p. 354–359. 134

Yamamoto, K., T. Kudo, A. Konagaya et Y. Matsumoto. 2003, «Protein name tagging for biomedical annotation in text», dans *the ACL 2003 workshop*, Association for Computational Linguistics, Morristown, NJ, USA, p. 65–72. 41

Zadeh, L., C. Negoita et H. Zimmermann. 1978, «Fuzzy sets as a basis for a theory of possibility», *Fuzzy sets and systems*, vol. 1, n o 3-28, p. 61–72. 55

Zerhouni, E. 2003, «The NIH Roadmap», *Science*, vol. 302, n o 5642, p. 63–72. 5

Zhong, M. et X. Huang. 2006, «Concept-based biomedical text retrieval», dans *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, NY, USA, p. 723–724. 56

Zhou, G. et J. Su. 2001, «Named entity recognition using an HMM-based chunk tagger», dans *the 40th Annual Meeting*, Association for Computational Linguistics, Morristown, NJ, USA, p. 473–480. 42

Zhou, W., T. Y. Clement, V. I. Torvik et N. R. Smalheiser. 2006, «A Concept-Based Framework for Passage Retrieval at Genomics.», *NIST Special Publication*, vol. 8. 57

Zhou, W., C. Yu, N. Smalheiser, V. Torvik et J. Hong. 2007, «Knowledge-intensive conceptual retrieval and passage extraction of biomedical literature», dans *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, New York, New York, USA, p. 655–662. 56, 57

Zhu, D. et B. Carterette. 2012, «Combining multi-level evidence for medical record retrieval», dans *Proceedings of the 2012 international workshop on Smart health and wellbeing*, ACM, New York, NY, USA, p. 49–56. 47