**École Doctorale Sciences Physiques,
Mathématiques et de l'Information pour l'Ingénieur
U.F.R DES SCIENCES ET TECHNIQUES**

**Centre de Recherche Universitaire
de Djibouti (CRUD)**

Recherche d'Information Multi Terminologique au sein d'un Dossier Patient Informatisé

THÈSE

présentée et soutenue publiquement le 22 Février 2012

pour l'obtention du

Doctorat de l'Université de Rouen
(***Spécialité : Informatique***)

Par

Ahmed-Diouf DIRIEH DIBAD

Composition du jury :

Directeur de thèse :	Stéfan Jacques Darmoni
Co-encadrant :	Philippe Massari
Président :	Thierry Lecroq
Rapporteurs :	Anita Burgun Marie-Christine Jaulent
Examinateurs :	Marc Cuggia Lina Fatima Soualmia
Invités :	Guillaume Savoye Jean-Philippe Leroy

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS EA 4108)

Á mes parents
Á toute ma famille
Á Hibo

REMERCIEMENTS

Au terme de ce doctorat, je tiens tout d'abord à remercier mon *DIEU le tout Puissant*, tout simplement pour m'avoir offert cette vie et cette opportunité.

Je souhaiterais, en premier lieu, exprimer ma profonde reconnaissance à Monsieur Stéfan Jacques Darmoni, qui a été, pour moi un très bon directeur de thèse. Je le remercie tout particulièrement de la confiance qu'il m'a témoignée, sa grande disponibilité et sa bonne humeur durant la réunion hebdomadaire de l'équipe.

Je tiens à remercier tout particulièrement, et à témoigner ma reconnaissance à Monsieur Philippe Massari, pour nos nombreuses discussions tant scientifiques que médicales, sa disponibilité tout au long de ce travail ; je le remercie également pour sa relecture constructive et approfondie de mon mémoire.

Je remercie Madame Marie-Christine Jaulent et Madame Anita Burgun qui ont accepté d'être mes rapporteurs et qui ont évalué mon travail. Je tiens à remercier trés sincèrement l'ensemble des membres du jury (Thierry Lecroq, Marc Cuggia, Lina F. Soualmia, Guillaume Savoye, Jean-Philippe Leroy,) qui me font le grand honneur d'avoir accepté de juger mon travail.

Je dédie cette thèse à mes parents, les mots me manquent et je ne pourrais résumer en quelques lignes toute ma reconnaissance pour eux. Je les remercie tendrement ainsi que ma famille (Fathiya, Ayane, Doubad, Kadidja, Fatouma, Bachir et Moustapha) pour leur soutien constant durant cette période...

Je ne manquerais au passage de remercier un exceptionnelle personne avec un fort tempérament qui a toujours été à côté de moi.

Il m'est impossible d'oublier de remercier nos institutions djiboutiennes à travers l'université de Djibouti et l'actuel président Monsieur Djama Moohamed Hassan ainsi que

son prédécesseur Monsieur Abdillahi Omar Bouh, qui soutiennnent inlassablement les enseignants dans leurs voeux d'entamer une carrière dans la recherche.

Je remercie le laboratoire LITIS EA 4108 de m'avoir accueilli en son sein et de m'avoir offert d'excellentes conditions de recherche. J'adresse tout particulièrement un grand merci chaleureux à tous les membres de l'équipe CISMeF (Badisse, Benoit, Catherine, Gaétan, Ivan, Romain, Sandrine, Tayeb) les cismefisés (Lina, Pereira, Josette), tous les docteurs qui m'ont précédé et même "dépassé", les doctorants (Zeid, Weim, Nicolas, Julien, Adila), leur souhaitant un très bon marathon... et sans oublier l'équipe TIBS.

Je tiens à remercier Christian Kala-Lobé pour son assistance lors de la mise en place des outils sémantiques d'Oracle.

Pour reprendre un ordre plus chronologique, je souhaite évidemment remercier mes amis de longue date, Mohamed Kileh Wais, pour sa disponibilité et en lui souhaitant une bonne fin de thèse, Ilyass Souleiman, Abdirahman M. Ahmed, Said Dabar Ouffaneh qui ont su me donner le courage et l'énergie de me relever à tout moment pour finir cette thèse *marathon*.

Mes plus vifs remerciements à mon ancien directeur, Mohamed Hassan Doualeh pour son esprit ouvert et pour m'avoir permis d'évoluer dans ma vie professionnelle et de m'orienter dans ce domaine de recherche.

Une spéciale dédicace aux membres du G6, Fatya, Hala, Samira, Chouaibe, Fanan (et A3D). Merci pour votre amitié et pour tous les moments de détente et fous rires inoubliables passés ensemble.

Enfin merci à toutes les personnes que je n'ai pas citées ici et qui se reconnaîtront dans ces quelques lignes.

Bien entendu, ce travail ne serait pas ce qu'il est devenu sans les discussions avec les chercheurs qui ont croisé mon chemin au cours des différentes rencontres scientifiques, sans leurs écrits qui ont été pour moi autant de chemins de lecture et qui ont apporté une base à cet édifice.

RESUME

Les progrès technologiques ont engendré de nouvelles possibilités d'accès et de traitement de l'information dans le domaine médical. Le dossier patient informatisé (DPI), dans cette optique, est considéré comme un outil de base pour construire des systèmes d'aide à la décision médicale (SADM). L'utilisation des données médicales et l'accès à une information concise sont devenus des enjeux majeurs pour les soignants. Plusieurs terminologies de références ont été créées à cet effet, et à des fins différentes, comme par exemple : les classifications CIM-10 et CCAM pour le codage épidémiologique et médico-économique, la nomenclature SNOMED pour le codage clinique, la classification ATC pour le codage des prescriptions médicamenteuses, et enfin la nomenclature LOINC pour le codage des analyses biologiques...Or, le constat est là : augmentation du volume des données médicales, données de plus en plus complexes difficilement accessibles, une grande partie de l'information est faiblement structurée et pour l'essentiel au format textuel, surplus d'informations... Devant ce constat et la nécessité grandissante de faire de la recherche d'information (RI) et d'exploiter les données médicales, il apparaît nécessaire d'intégrer au sein du DPI, outre les fonctions de gestion et de communication, des fonctions de RI. Ceci nécessite de développer des outils de RI proches des outils existants pour la RI documentaire. Ces outils de RI orientés DPI se fondent sur des modèles d'information [de données] permettant de faire de la RI et indépendemment des modèles du système de gestion du DPI. Notre contribution consiste principalement à concevoir un modèle de données générique adapté à la RI. Ce modèle, fondé sur la notion d'éléments informationnels du DPI, est conçu dans une approche hybride (approche Entité-Relation, approche fondée sur les métadonnées) pour avoir un modèle multi niveau afin de réaliser l'interopérabilité, non seulement au niveau des données mais aussi au niveau sémantique, et de concilier un niveau de granularité pour une meilleure scalabilité. Notre proposition a été validée par le développement du prototype RIDoPI et par une *preuve de concept* avec les technologies du Web Sémantique.

Mots-clés : Dossier médical informatisé, Recherche et stockage d'information, Système d'Information, Web Sémantique, Modélisation de données.

ABSTRACT

Recent technological advances have led to new opportunities for the access and processing of data in medicine. The Electronic Health Record (EHR), in this context, is considered to be a basic tool for building a Clinical Decision Support System (CDSS). The use of medical data and the access to concise information have become of major importance for clinicians. Numerous medical terminologies have been developed to meet different purposes : ICD-10 and CCAM classifications for epidemiologic, medico-economic coding, SNOMED nomenclature for clinical coding, ATC classification for drug prescriptions coding, and finally LOINC nomenclature for biological coding analysis. However, the fact is : that there is an increasing of amount of data, which has become more complex with difficult access. Also much of the information is poorly structured and mainly in a text format, including an information overload. Subsequently, as there is a growing need to perform information retrieval (IR) of medical data, it will be necessary to include in EHRs, in addition to their management and communication functions, as well as IR functions. This requires the development of IR tools close to existing Web IR tools. The IR tools oriented by EHR are based on information model [data] for IR, and independent from EHR management system models. The aim of our research was to mainly develop a generic data model adapted to IR. This model, based on information elements concept of EHR, is designed in a hybrid approach (Entity-Relationship approach, metadata approach) for a multi-level model in order to achieve an interoperability, not only on the data level but also on the semantic level, and to combine a level of granularity for better scalability. Based on our research, we were able to develop a RIDoPI prototype using a *proof of concept* with Semantic Web technologies.

Key-words : Electronic health records, Information storage and retrieval, Information system, Semantic Web, Data Modeling.

PREAMBULE

Il est bien admis par tous que l'utilisation des outils TIC en santé est le seul moyen pour organiser la santé de demain.
L'*e*-Santé(*e*-Health), un paradigme d'application des TIC au domaine de santé s'est développé parallèlement aux progrès des technologies des réseaux et de gestion de l'information. Son objectif premier est l'amélioration de la qualité des soins d'un patient dans cette société civile globale avec en toile de fond l'atteinte des Objectifs du Millénaire pour le Développement (OMD) d'ici 2015.
Les approches méthodologiques et les stratégies de développement, pour parvenir à ces objectifs, diffèrent d'un contexte à l'autre, d'un pays à l'autre.
Les pays développés ont mis en place, pour chaque structure de santé, un SI pour faciliter la collecte, la saisie, le stockage et la recherche des données cliniques du dossier patient informatisé jadis effectués manuellement dans les documents médicaux papiers. Passées les étapes d'informatisation du domaine, ces pays ont fait face à l'épineux problème d'interopérabilité *technique* et *sémantique* de ces systèmes... Problémantiques (quasiment) résolues, dans une approche transversale avec une large couverture du domaine, par la mise en place de vocabulaires et de normes d'échange standards... Les nouvelles pratiques de la médecine (la télémédecine, les soins à domicile, les services personnalisés de soins, la médecine translationnelle, ...) ont affecté le développement de ces SI en les obligeant à adopter d'autres approches.
Qu'en est-il dans les pays en voie de développement ? Des contextes spécifiques : développer des outils pour un domaine médical spécifique (épidémiologie du SIDA) plutôt qu'une informatisation globale des processus de soins. Des stratégies locales ne prenant en compte qu'une dimension du domaine hospitalier : faciliter la collecte fiable et facile des données à des fins d'analyses à visée épidémiologique ou de planification de santé à travers des projets pilotes d'implémentation de SI communément appelés en anglais District Health Information System (DHIS). L'hétérogénéité de ces systèmes, la non disponibilité [1] des données et d'outils d'aide à la décision, au moment opportun, pour

1. Mes travaux rentrent dans le cadre d'une cotutelle entre l'université de Rouen et l'université de Djibouti. Par conséquent, cette non disponibilité de modèles de SI est aussi une cause pour la non-

décider efficacement, sont des manques pour répondre aux besoins des OMD. Ces SI, qui deviennent pervasifs nécessitant des recherches sur l'accès et la présentation des données, peuvent-ils contribuer à combler ces lacunes ? D'autant plus que ces pays se retrouvent avec des technologies qu'ils ne sont pas techniquement en mesure de maintenir et qui n'ont pas suffisamment de personnels pour les réutiliser.
Finalement, aucune approche ne peut réussir sans une analyse détaillée globale et spécifique des besoins : *globale* pour ne pas se restreindre à un sous domaine particulier du domaine de santé et *spécifique* pour s'adapter au contexte local du pays.... Par conséquent, pour arriver à une infrastructure `génératrice d'économies`, la tendance s'est orientée vers des SI `à la carte` avec des approches de développement orientées services (SOA) qui font passer la vision `traditionnelle` des SI d'outils de stockage de données à des composants des réseaux socio-économiques. Est-ce la solution ?
Ceci dit, le domaine de santé est un socle de recherche et un domaine encore peu exploré (une accentuation des travaux durant ces dix dernières années) par les chercheurs. Les recherches en SI représentent un sujet de recherche trés *vaste* mais aussi très varié voire *complexe* à mener à bien. *Vaste* car le champ d'action s'étend de la structuration des bases de données jusqu'à la mise en place d'outils d'exploitation des données parmi lesquels des applications de recherche d'information pour faciliter l'accès à l'information pertinente au moment opportun et de n'importe quel lieu. *Complexe* car les données sont hétérogènes et stockées dans des SGBDR dans des formats variés (pour la plupart au format texte). Et c'est pour cela que notre contribution scientifique se situe au niveau de la représentation intelligente de ces données sous une forme exploitable plus pratique pour permettre leur (ré)utilisation par des systèmes informatiques.

---

application de mes travaux aux données patients de mon pays. Cependant, un plan quinquennal (2012-2017) va être mis en place pour le *premier et futur* systèmes d'information de santé à l'échelle regionale et nationale, et la création du *premier et futur CHU* **Hopital Peltier** de la République de Djibouti.

Table des matières

Table des figures

Liste des tableaux

Liste des Abbréviations

ANAES Agence Nationale d'Accréditation et d'Evaluation en Santé

ANDEM Agence Nationale pour le Développement de l'Évaluation Médicale

ANR Agence Nationale de la Recherche

BDO Bases de données-objets

BDR Bases de données relationnelles

CDA Clinical Document Architecture

CDP C-PAGE Dossier Patient

CLEF Clinical e-Science Framework

CPOE Computerized Physician Order Entry

DHIS District Health Information System

DMP Dossier Médical Personnel

DPI Dossier Patient Informatisé

DW Datawarehouse

EIDM Elements Informationels du Dossier Médical

ETL Extract Load Transformation

F-MTI French Multi-Terminology Indexer

HAS Haute Autorité de Santé

HeTOP Health Terminology Ontology Poral

HIMSS Healthcare Information and Management Systems Society

HL7 Health Level Seven

HMORN Health Maintenance Organization Research Network

IA Intelligence Artificielle

InterSTIS Interopérabilité Sémantique des Terminologies dans les Systèmes d'Information de Santé français

IPAQSS Indicateurs Pour l'Amélioration de la Qualité et de la Sécurité des Soins

MAIF MeSH Automatic Indexing in French

NCBI National for Center Biotechnology Information - Computational Biology Branch

OMD Objectifs du Millénaire pour le Développement

OMOP Observational Medical Outomes Partnership

PACS Picture Archiving and Communication System (système d'archivage et de transmission d'images)

PIM Portail d'Information sur le Médicament

PLAIR Plateforme d'Indexation Régionale

PMSI Programme de Médicalisation du Système d'Information

PSIP Patient Safety Through Intelligent Procedures in Medication

PTS Portail Terminologique de Santé

RAVEL Recherche et Visualisation des informations dans le dossier patient électronique

RIDoPI Recherche d'Information dans le DOssier Patient Informatisé

SADM Systèmes d'Aide à la Décision Médicale

SED Système d'Entrepôt de Données

SGBDR Système de Gestion des Bases de Données Relationnelles

SGL Système de Gestion de Laboratoire

SI Système d'Information

SIC Systèmes d'Information Clinique

SIDA Syndrome de l'ImmunoDéficience Acquise

SIH Systèmes d'Information Hospitaliers

SIS Systèmes d'Information de Santé

SMK Structured Meta Knowledge

SRI Système de Recherch d'Information

T2A Tarification à l'Activité

TD*IDF Term Frequency- Inverse Document Frequency

TecSan Technologie pour la Santé et l'Autonomie

TerSan Terminologies d'Interfaces en Santé

TIBS Traitement de l'Information en Biologie et Santé

TIC Technologies de l'Information et des Communications

UMLS Unified Medical Language System

VCM Visualisation de Connaissances Médicales

INTRODUCTION GENERALE

[Hersh, 2008] a classifié les informations du domaine médical, en 2 catégories :

- les informations spécifiques au patient ou aux informations de soins (dossier patient) ;
- les informations spécifiques aux connaissances du domaine issues des recherches expérimentales et observationnelles.

La demande de l'information a toujours été et restera toujours essentielle pour les professionnels, autant *pour accéder à l'histoire de santé* du patient afin de prendre une décision (dans le cas précis du processus de prise en charge du patient), que *pour accéder aux connaissances* issues des expériences et des observations du domaine, et de les appliquer dans ce processus décisionnel. Pour répondre à cette demande d'informations spécifiques au patient, l'informatisation du dossier de santé, par le biais des outils TIC, avait pour objectifs, pour n'en citer que ces deux éléments importants :

- la *mémorisation* de l'information dans un format de stockage (majoritairement dans des SGBDR et des SED) selon un modèle sémantique de données pour faciliter son utilisabilité ;
- l'*exploitation* de l'information à travers des outils d'aide à la décision , dont le seul but, n'est autre que l'accès à l'information stockée, dans des délais compatibles aux pratiques cliniques.

Les systèmes d'aide à la décision peuvent pallier ces difficultés, mais la plupart de leurs modalités d'intervention, recensées par [Renaud-Salis et al., 2010], nécessitent de trouver les informations nécessaires dans les dossiers médicaux. Des outils efficaces orientés *SADM* ont été développés : *Les systèmes de recherche d'information* . Les SRI gèrent un volume d'informations relativement considérable et offrent des méthodologies de recherche et de traitement permettant de retrouver l'information en lien avec un besoin informationnel [Sakji, 2010]. Dans ce contexte, la recherche d'information au sein du dossier patient informatisé [Dirieh Dibad et al., 2009], pourrait devenir une activité quo-

tidienne largement pratiquée par les soignants telle que la recherche d'information sur le Web effectuée via des moteurs de recherche généralistes (comme Google) ou plus spécialisés (catalogue CISMeF) [Darmoni et al., 2001b].
Les SRI manipulent, dans le cas du dossier patient informatisé, une collection de données traduisant, d'une part une hétérogeneité des données médicales qu'il convient de les unifier à travers un modèle sémantique de données, d'autre part une sémantique implicite à découvrir à l'aide d'outils et de méthodes de recherche et de traitement.
Á ce niveau, la question qui se pose est alors : *Face à l'évolution de la médecine qui a conduit à une augmentation de la quantité et la complexité des données médicales, est-ce que les soignants ont à leur disposition les outils informatiques,qui pourront répondre à leurs besoins informationnels ?.*

Problématiques et Objectifs

Problématiques médicales

Les questions posées par un professionel peuvent être de différentes natures : question sur l'histoire de la maladie du patient (*Quel sont les traitements médicamenteux prescrits au patient pour son diabète ?*), questions sur la prise en charge d'une pathologie, d'un symptôme (*Quelles sont les recommandations pour la prise en charge d'une adénopathie cervicale ?*)... et peuvent être très complexes, des questions sur les patients qui répondent à des critères d'inclusion/d'exclusion pour des essais thérapeutiques.
Des aides existent pour affiner ses questions sur les catalogues de : recommandations francophones professionnelles [2], Vidal Recos [3], ... mais plus de 50% des questions n'ont de réponses qu'au sein du dossier patient informatisé et sont difficilement accessibles [Osheroff et al., 1991].
La disponibilité du dossier patient informatisé n'est pas une condition suffisante pour améliorer la qualité des soins [Schumacher et al., 2010]. Au délà des vues proposées par le DPI [Zeng et al., 2002; Massari et al., 2008] pour faciliter l'accès à l'information, il doit intégrer des fonctions clefs comme les SRI. D'un point de vue médical, notre objectif est simple : "accéder aux connaissances contenues dans ce gisement de données que contient le dossier patient informatisé". Cela nécessite le développement de prototypes d'interface de RI permettant de rechercher aussi bien dans un seul dossier que dans une base de dossiers médicaux.

Problématiques théoriques

La RI dans le domaine de soin est un domaine nouveau [Hersh, 2008], et en train de se développer en mettant en oeuvre les outils et les méthodes courantes déjà implémentés dans la RI documentaire [Sakji, 2010]. Plusieurs travaux se sont intéressés à cette problématique en développant des outils dédiés à la RI selon des approches différentes et à des besoins différents : [Deshmukh et al., 2009; Cuggia et al., 2010] pour la recherche translationnelle, [Currie et al., 2001] pour identifier les patients qui ont des problèmes de coeur et qui fument. [Joubert et al., 1996] proposent une naviguation du dossier patient à travers le métathésaurus UMLS dans le cadre du projet ARIANE [4] et enfin [Chute and Yang, 1992] à travers l'outil SMART pour rechercher des rapports de chirurgie.
Parmi les grandes questions courantes à ce jour dans les travaux du domaine de la recherche d'information au sein du dossier patient informatisé, deux questions revêtent

2. CISMeF Bonnes Pratiques, URL : http ://doccismef.chu-rouen.fr/servlets/CISMeFBP
3. L'essentiel sur les recommandations thérapeutique, URL : http ://www.vidalrecos.fr
4. Projet ARIANE, URL : http ://cybertim.timone.univ-mrs.fr/recherche/projets-recherche/ARIANE

une importance déterminante, d'impact important sur l'efficacité des SRI : *modèles - recherche d'information.*
La première question traite des problématiques inhérentes à la modélisation des données du dossier patient informatisé. Cette modélisation cible deux principaux objectifs : (1) disposer d'un modèle générique et flexible pour faciliter l'intégration de l'ensemble des données pertinentes au besoin informationnel des soignants, et (2) réaliser l'interopérabilité, non seulement au niveau des données mais aussi au niveau de la sémantique afin de faciliter l'intégration de nouvelles connaissances. De ce point de vue, deux éléments entrent en jeu : (a) les terminologies médicales pour leur rôle dans la description de la sémantique des données, et (b) les normes standards de communication pour permettre la transposition du modèle dans d'autres systèmes (modèles).
Cette première question met, alors en amont, le besoin de disposer d'une information médicale sous une forme *exploitable* de manière informatique et *standardisée* qui est un domaine d'intérêt de la seconde question.

Objectifs

Dans ce travail, nous cherchons principalement à concevoir un modèle de données générique adapté à la RI au sein d'un dossier patient informatisé et dans un contexte multi terminologique (fondé sur plusieurs terminologies médicales). Nous présenterons le contexte du travail dans lequel s'est déroulée la thèse par une brève présentation de l'équipe CISMeF.

Contexte du travail

CISMeF est un groupe de recherche sur la gestion de la connaissance et les systèmes d'information de santé (GCSIS) [5], qui forme, avec un groupe de bio-informaticiens, l'équipe TIBS du laboratoire LITIS EA 4108. Au sein du GCSIS et sous la codirection du responsable des technologies de l'information et de la communication (Stéfan J. Darmoni) et du conservateur de la bibliothèque médicale (Benoît Thirion) du Centre Hospitalier Universitaire de Rouen, un projet prédomine CISMeF [6](acronyme de Catalogue et Index des Sites Médicaux Francophones) qui est devenu depuis 1995 le site catalogue de référence dans le monde francophone de la santé sur l'Internet [Darmoni et al., 2001b].
Entre 1995 et 2005, l'indexation des ressources Internet s'effectuait manuellement par une équipe de quatre documentalistes (en 2011). Cette indexation était fondée exclu-

5. http ://www.chu-rouen.fr/l@stics/
6. Catalogue et Index des Sites Médicaux Francophones, URL : www.cismef.org

sivement sur le thésaurus MeSH [7], tandis que la description des ressources s'appuyait sur un ensemble de métadonnées dont le Dublin Core [Darmoni et al., 2001a]. De nombreuses améliorations ont été apportées au thésaurus MeSH (types de ressources pour la description et l'indexation, métatermes pour la recherche d'information et la catégorisation, définitions en français, ajout de synonymes, affiliation des types de ressources, indexation majeur/mineur pour les types de ressources et les métatermes, ...) [Thirion et al., 2003]. Depuis 2007, l'équipe CISMeF s'intéresse à trois domaines de recherche, très interpénétrés :

- L'indexation mono-terminologique (autour du MeSH) des ressources Web avec l'outil MAIF développé pendant la thèse d'Aurélie Névéol [Névéol, 2005], actuellement en post-doc [8] à la NLM dans la branche CBB de NCBI . Cette indexation est maintenant multi-terminologique avec l'outil F-MTI d'indexation automatique multi-terminologique, multi-documents et multi-tâches capable de produire une proposition d'indexation pour les documents de santé, pendant la thèse de Suzanne Pereira [Pereira, 2008], actuellement Responsable Projet R&D chez VIDAL ;
- Le développement d'outils et de méthodes d'alignement de terminologies et d'ontologies médicales fondés sur des méthodes TAL et sémantique, pendant la thèse de Tayeb Merabti [Merabti, 2010] sur l'interopérabilité sémantique intra et inter terminologies de santé. Il est acutellement en post-doc dans l'équipe dans le cadre du projet PLAIR (Plateforme d'Indexation Régionale) [9].
- La recherche d'information mono-terminologique, étudiée pendant la thèse de Lina Soualmia [Soualmia, 2004], actuellement maître de conférences (27ème section) au LIM&BIO [10] (Paris XIII). Pour faciliter la tâche des utilisateurs, une recherche d'information implicite a été mise en oeuvre avec le système KnewQuE (Knewledge-based Query Expansion) afin de corriger, préciser et enrichir les requêtes des utilisateurs. Cette recherche d'information est maintenant multi-terminologique au sein du catalogue CISMeF, notamment sur les médicaments (codes ATC, CIP, CIS, UCD) à travers le portail bilingue PIM développé pendant la thèse de Saoussen Sakji [Sakji, 2010].

De nombreux projets en collaboration avec des industriels et des équipes académiques ont vu le jour parmi les principaux :

7. US National Library of Medicine - Medical Subject Headings, URL : http ://www.nlm.nih.gov/mesh/meshhome.html
8. NCBI Postdoctoral Fellowship Program, URL : http ://www.ncbi.nlm.nih.gov/Sitemap/Summary/postdoc.ht
9. URL : http ://www.plair.org
10. http ://www-limbio.smbh.univ-paris13.fr/site/index.php

- Le projet L3IM [11](Langage Iconique et Interfaces Interactives en Médecine) qui a pour finalité d'offrir un accès rapide à des informations médicales. Cette approche est rendue possible grâce au langage iconique VCM qui permet de représenter un ensemble de concepts médicaux comme des maladies, des médicaments ou encore des examens complémentaires [Lamy et al., 2010] ;
- Le projet PSIP qui a pour objectif de développer des systèms d'alertes sur la détection d'effets indésidérables aux médicaments à travers des bases de connaissances [Ammenwerth et al., 2011]. L'équipe CISMeF a été en charge du *semantic mining* et de la création du Portail Terminologique de Santé (PTS) [12] ;
- Le projet InterSTIS [13] qui a pour but de rendre interopérables les principales terminologies médicales.

Dans un cadre plus large, l'équipe CISMeF a développé, à partir de 2011, un prototype de portail multi-terminologique et inter-lingue HeTOP (Health Terminology Ontology Poral) [Grosjean et al., 2011b].
Nos travaux trouvent leur `point zéro` aux changements de stratégies de l'équipe CISMeF pour adapter ses méthodes et outils existants (en particulier le moteur de recherche), à l'exploitation des données du dossier patient informatisé en s'appuyant initialement sur le SIC du CHU de Rouen.

11. URL : http ://projet4-limbio.smbh.univ-paris13.fr/
12. URL : http ://pts.chu-rouen.fr/
13. URL : http ://www.interstis.org/

Organisation du mémoire

Le mémoire retraçant nos travaux, est composé de 6 chapitres princpaux s'articulant en deux parties selon le schéma synoptique défini ci-après (cf. Figure 0.1).
La **première partie** traite du dossier de santé, la modélisation et de la recherche d'information.
Le premier chapitre résume les différentes étapes d'informatisation du dossier de santé pour répondre à cette demande de l'information.
Le `second chapitre` présente la problématique de la modélisation des données du dossier patient ainsi que les différents modèles proposés dans la littérature. Inclue dans ce chapitre, la présentation des modèles de représentation des données et des connaissances pour une interopérabilite *technique* et *sémantique* des SIS.
Le `troisième chapitre` traite de la problématique de la recherche d'information inhérentes au contexte de soins. Inclue dans ce chapitre, une comparaison entre le prototype **RIDoPI** et les différents SRI existants dans la littérature.
La **deuxième partie** présente notre contribution à la mise en oeuvre de stratégies de recherche d'information dans le DPI à travers notre approche méthodologique pour la conception d'un modèle de données générique adapté à cette RI.
Le `quatrième chapitre` détaille globalemet notre approche. Nous y décrivons notamment les contraintes de modélisation, une évaluation de certains modèles proposés dans la littérature pour expliquer leur non-application dans le cadre de ces travaux, le modèle EI@DM que nous proposons. Enfin, une analyse comparative (non exhaustive) des modèles évalués (incluant le modèle EI@DM) et une discussion critique pour le positionnement de ce travail.
Le `cinquième chapitre` présente deux études de cas pour une expérimentation du modèle EI@DM. Nous détaillons les fondamentaux des architectures implémentées.
Le `sixième chapitre` décrit les résultats d'expérimentation pour une validation applicative du modèle EI@DM. Cette évaluation a pour but de démontrer l'adaptation du modèle à la RI dans les deux études de cas. Nous y présentons aussi une discussion critique de l'évaluation, comme cadre de refléxion pour définir les nouvelles directions de nos travaux.
En conclusion, nous dressons un bilan de nos travaux, en mettant en exergue les éléments importants. En plus, nous présentons ensuite les perspectives d'évolution des travaux.

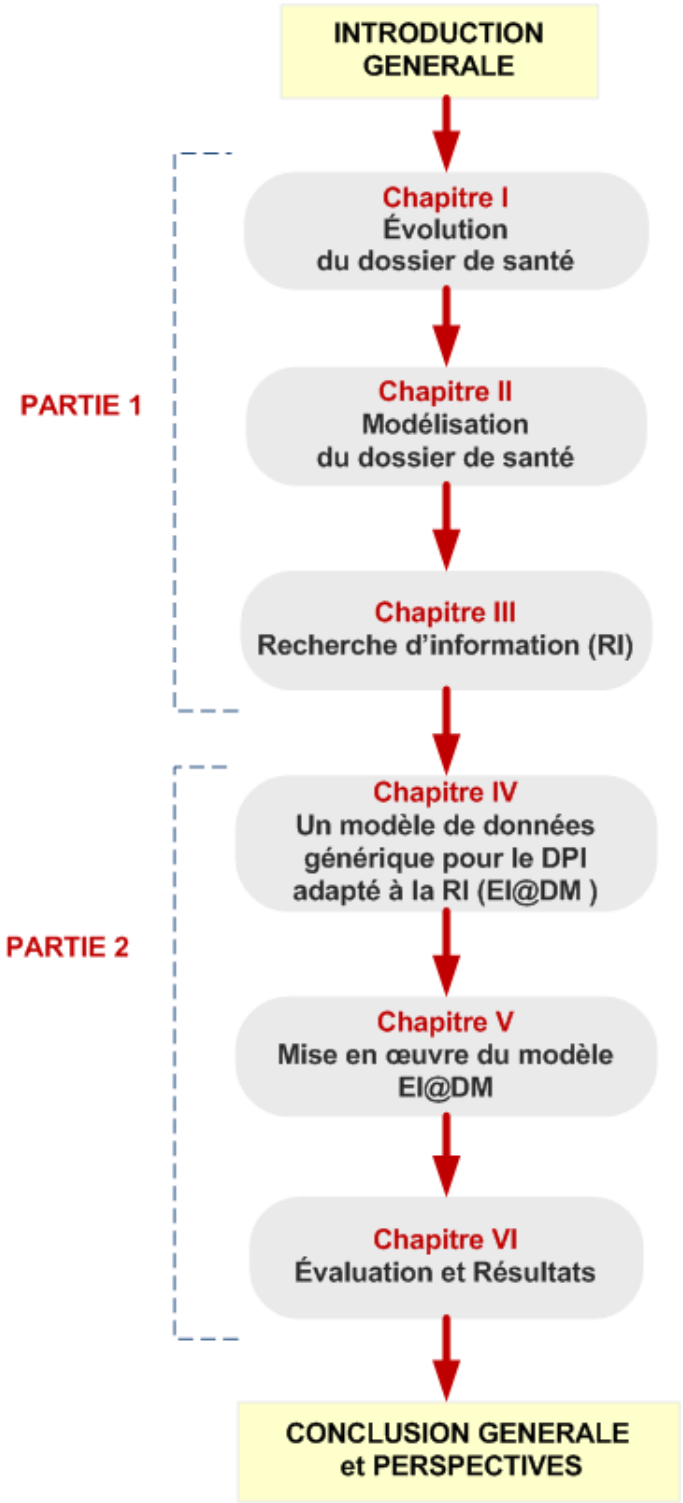

FIGURE 0.1 – Schéma synoptique du manuscrit

Première partie

Dossier médical, Modélisation et Recherche d'Information

CHAPITRE 1

Le dossier de santé

Introduction

La démarche médicale est fondée sur l'observation du malade et la production d'informations. Depuis des siècles, ces informations ont été rédigées par les médecins sur des registres. Longtemps en effet, l'observation médicale ne servit pas aux soins, la mémoire du médecin, seul à s'occuper du malade et disposant de moyens thérapeutiques limités, était suffisante. Ces observations ont pris, au 19ième siècle, une forme qu'on connait aujourd'hui : le dossier médical [1] pour être la mémoire médicale des PDS [Moutel, 2009]. Le partenariat, autour d'un même patient, des PDS pour collaborer et coopérer afin de mieux dialoguer avec le patient, la place du dossier médical dans le système de santé comme un outil d'évaluation de la qualité des soins et des pratiques médicales, ont largement modifié les différents modes d'exercice de la médecine(médecine de ville, médecine hospitalière, médecine à domicile) et appellent à la mise en place de systèmes de communication adaptés [Fieschi, 2003]. Ainsi donc les progrès de la médecine, des TIC et des droits des patients ont engendré une nécessaire inéluctable évolution du dossier médical [Haux et al., 2002; Hasman et al., 2003].
Dans ce chapitre, nous allons décrire brièvement l'évolution au fil du temps du dossier médical avant d'aborder son informatisation et *a posteriori* de l'informatisation du système de santé qui passe par une mutation rapide de l'informatique hospitalière et extrahospitalière. Notre approche globale est résumée dans la figure 1.1.

1. Article R1112-2 du code de la santé publique

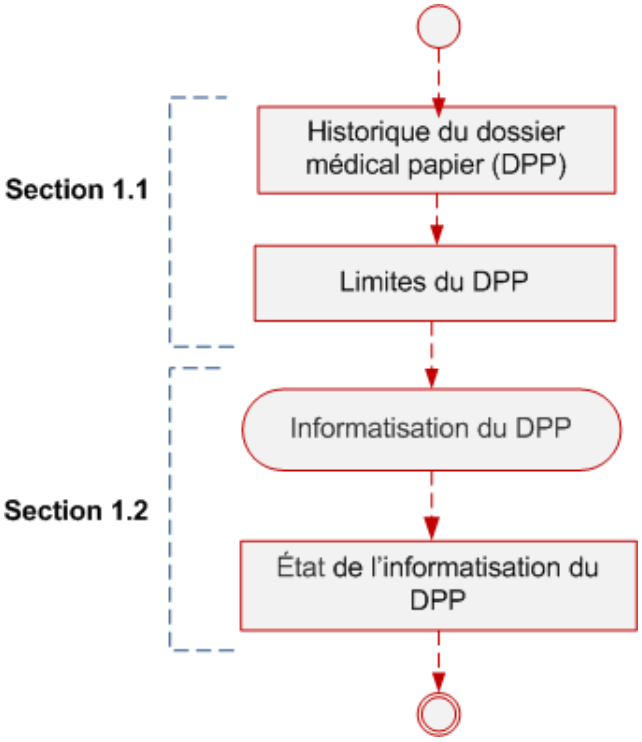

FIGURE 1.1 – Schéma synoptique de l'évolution du dossier de santé

1.1 Le dossier médical papier : des origines à nos jours

Les évolutions du dossier médical ont, jusqu'au milieu du 20e siècle, été induites par les évolutions des pratiques médicales. Plus récemment ces évolutions ont été liées au évolutions organisationnelles et à la mise en place d'une législation et de textes réglementaires en précisant en particulier : le contenu, la durée d'archivage et des modalités de gestion. Nous scinderons cet historique en 2 parties :

- Le dossier médical avant 1950 ;
- Le dossier médical après 1950.

Pour cette seconde période nous décrirons l'organisation du système de santé français et les évolutions légales et réglementaires ayant eu un impact sur le dossier médical.

1.1.1 Le dossier médical avant 1950

La mémoire du médecin était autrefois suffisante pour enregistrer les données nécessaires au suivi du patient, d'autant plus que dans la quasi totalité des cas un seul médecin le prenait en charge. Les données médicales étaient rassemblées dans des registres à visée épidémiologique, nosologique et d'enseignement [Moutel, 2009].
Pour la plupart des auteurs :

- Le dossier clinique, à l'époque d'Hippocrate et jusqu'au début du 17ème siècle était essentiellement considéré comme un outil d'enseignement sur l'art de prendre des observations pour en tirer des comptes-rendus exploitables, mais aussi comme un outil pour consigner les facteurs risques et les causes d'une maladie [Van Bemmel and Musen, 1997] ;
- Il faut remonter au 5ème siècle, pour trouver les premières traces du dossier médical. Á cette époque, les médecins arabes (tels que Rhazès (865-925), Avicenne (930-1037) ou Avenzoar (1073-1162)), créent la médecine clinique en consignant sur leurs livres les observations médicales des cas intéressants([Sournia, 1991; Callebat, 1999; Moutel, 2009]).

Quant aux médecins égyptiens, ils prennent des notes et consignent leurs observations sur des rouleaux de papyrus. Un des papyrus a été retrouvé en 1874 par G. EBERTH à Memphis, il contenait notamment plus de 800 prescriptions et d'autres données médicales [Feldman and Goodrich, 1999]. De nombreux écrits concernant des cas cliniques antérieurs au 5ıème siècle ont été retrouvés, il est difficile de déterminer s'ils n'étaient que des descriptions de cas à visée épidémiologique, nosologique ou d'enseignement, ou

les prémisses de dossier médicaux. Cet extrait d'un texte de Gerbert datant du 9ème siècle évoque déjà l'utilité du dossier médical dans le suivi d'un patient :
"...En ce qui concerne la situation du frère souffrant de la maladie de la pierre, je pourrais m'en occuper si je pouvais savoir ce qui a été prescrit par les autorités médicales. Actuellement contentes toi d'une petite dose de l'antidote philantropos et de la recette qui l'accompagne et ne t'en prends qu'à toi même si, en ne suivant pas les prescriptions, tu utilises mal ce qui est prévu pour rendre la santé" [Riché, 1979].

Ce n'est réellement qu'au début du 19ième siècle que le concept de dossier médical attaché à chaque patient comme nous le connaissons apparaîtra en France dans les Hôtels-Dieu [Moutel, 2009] mais son contenu reste succinct : il inclut des données médicales, sociales et administratives. La fulgurante avancée de la recherche médicale et des progrès de la médecine, la nécessité de mieux suivre le patient dans sa prise en charge, ont inéluctablement amené le dossier médical à évoluer. Le rapport de [Flexner, 1910] sur la formation médicale est la première déclaration officielle faite sur le contenu et les fonctions du dossier médical. En 1907 à la Mayo Clinic, le Dr Henry S. Plummer met en place un dossier médical qui peut être considéré comme le premier dossier médial hospitalier "moderne"*(medical record)*. Un dossier est créé pour chaque patient à qui il est attribué un numéro lors de sa première venue dans l'établissement, ce numéro et ce dossier seront utilisés lors des prises en charge suivantes. Ce numéro unique de la Mayo Clinic est le précurseur de l'Identifiant Patient Permanent(IPP) de nos SIH . Outre la volonté d'avoir un meilleur recueil d'informations par l'intermédiaire de formulaires types, une des raisons majeure de la mise en place de ces dossiers est la prise en charge des patients par une équipe de médecins et de chirurgiens.

Ainsi à l'aube des grands tournants en ce qui concerne les droits des patients [CNEH, 2011] et de grandes mutations organisationnelles en particulier de la médecine hospitalière, la rédaction des informations recueillies et des actions effectuées dans le cadre du soin dans un document propre à chaque patient qui se généralise du moins dans certains hôpitaux. Jusqu'au milieu du 20ième siècle, son contenu était limité ainsi que son partage, il s'agit du dossier d'un médecin, au mieux d'un service.

1.1.2 Le dossier médical après 1950

Le dossier médical au cours de cette période va prendre une importance considérable. Ses évolutions sont liées aux modifications des pratiques médicales, en particulier la prise en charge partagée des patients, mais aussi induites par les transformations de notre système de santé et en ce qui concerne les dossiers hospitaliers une réglementation spécifique. Comme nous le verrons dans une brève description du système de santé français, les différentes pratiques et réglementations de la médecine hospitalière et de la médecine libérale ambulatoire nécessitent de différencier la description des dossiers de médecine de ville des dossiers hospitaliers.

1.1.2.1 Le système de santé français et ses différentes évolutions

Au cours de la deuxième moitié du 20ième siècle, le système de santé français subit d'importantes mutations en particulier en ce qui concerne les hôpitaux, nous ne mentionnerons que celles qui ont eu un impact sur le dossier médical. Les centres hospitaliers régionaux, le découpage en services, la commission médicale consultative sont nés (Loi du 21 décembre 1941 *"dite Charte hospitalière"*, décret du 17 avril 1943) [2].
Le système hospitalier est profondément modifié par les ordonnances de 1958 et la Loi nn° 70-1318 du 31 décembre 1970, il devient un élément majeur du système de santé, sa gestion est fortement encadrée par les organes décentralisés de l'état. Les capacités en nombre de lits, les équipements lourds des hôpitaux publics et privés sont définis par la carte sanitaire. Par contre aucune régulation de la médecine de cabinet n'est mise en place, sauvegardant la libre installation qui persiste jusqu'à maintenant.

Les autres évolutions qui vont impacter le dossier médical sont :
- les modes de financements des hôpitaux ;
- la mise en place de procédures qualité et de sécurité sanitaire ;.
- l'incitation à la création de réseaux de soins ;
- l'accroissement des recours juridiques des patients ;

Le financement du système de santé français, au delà des 30 glorieuses a été une des préoccupations importantes des gouvernements successifs. Le financement des hôpitaux publics jusqu'en 1984 reposait sur le "prix de journée" ce qui revenait de fait à définir les recettes en fonction des dépenses de l'année précédente. Le financement de la médecine libérale ambulatoire a toujours reposé sur le paiement à l'acte des prestations. La loi nn° 83-25 du 19 janvier 1983 et son décret d'application n° 83-744 du 11 août 1983

2. Institut de recherche et documentation en économie de la santé (IRDES), URL : http ://www.irdes.fr/EspaceDoc/DossiersBiblios/HistoriqueReformesHospitalieres.pdf

mettent en oeuvre dans les hôpitaux publics la dotation globale de financement. Afin d'adapter cette dotation globale à l'activité, est progressivement mis en place au début des années 90 le projet puis le Programme de Médicalisation du Système d'Information (PMSI) s'inspirant des DRG américains. Il impose le codage des séjours hospitaliers en CIM9 puis CIM 10, le codage des actes (CDAM puis CCAM) et la transmission des données, en 1994 pour les hôpitaux publics, en 1995 pour les hôpitaux privés. Depuis 2005 les recettes hospitalières (publiques et privées) dépendent essentiellement de l'activité hospitalière mesurée selon un système national dénommé T2A.
Les démarches qualité et sécurité des soins se mettent en place progressivement depuis le début des années 80, une coordination nationale de ces démarches est initiée en 1990 par la création de l'ANDEM, à laquelle succède l'ANAES en 1996, ce qui marque le début de l'accréditation des établissements. La Haute Autorité de Santé (HAS), remplace l'ANAES en janvier 2005, avec des missions plus larges, la certification remplace l'accréditation, des IPQASS apparaissent. Le dossier médical hospitalier est fortement impacté par l'accréditation, puis la certification et les IPAQSS. La mise en place ou le renforcement des vigilances sanitaires imposent le signalement des incidents et le recueil des informations permettant la traçabilité (produits sanguins, dispositifs implantables...).
Un autre élément important dans l'évolution de notre système de santé est l'incitation des pouvoirs publics a une meilleure coopération sanitaire, par l'intermédiaire de syndicats inter hospitaliers (loi du 31 juillet 1991) puis de groupements de coopération sanitaire (ordonnance du 4 septembre 2003).
Enfin, au cours des dernières décennies le comportement des patients vis à vis de la médecine et des médecins se modifie[3] : *"Le patient contemporain entend non seulement être soigné - et bien soigné - ce qui est normal ; il veut, en outre, que le médecin le guérisse, ce qui est tout autre chose. Et bien souvent, si cette guérison n'est pas obtenue, le mécontentement du malade déçu se traduira par une mise en cause de la responsabilité professionnelle du praticien".*
Ainsi, en 2011 notre système de santé est composé d'établissements hospitaliers publics dont certains ont des capacités en nombre de lits importantes (en particulier les CHU), d'établissements hospitaliers privés et de cabinets de médecine libérale qui se partagent la prise en charge des patients. Le suivi du patient ne peut plus reposer sur la mémoire individuelle des médecins, ce qui rend indispensable le dossier patient qui outre les informations directement liées aux soins doit gérer celles nécessaires à la facturation, à la traçabilité et à l'évaluation de la qualité des soins et avoir une valeur probante devant la justice en cas de contentieu.

3. http ://www.ethique.inserm.fr/ethique/Ethique.nsf/397fe8563d75f39bc12563f60028ec43/b0cf353d60846663c1

1.1.2.1.1 Dossiers hospitaliers

D'un dossier minimaliste, le dossier médical hospitalier va progressivement se transformer en un outil d'archivage de tous les documents utilisés au cours des hospitalisations ou des consultations. Son volume va augmenter avec la durée de suivi des patients et la multiplication des documents liés au plus grand nombre d'investigations complémentaires pratiquées, aux nécessités de justification financière et de la qualité de la prise en charge, à la traçabilité et à la multiplication des documents pouvant être utiles en cas de contentieu (autorisation d'opérer, documents prouvant que le patient a reçu l'information nécessaire ...). Bien que le contenu du dossier médical hospitalier ne soit ni légalement ni réglementairement défini, il apparaît une première reconnaissance réglementaire sous forme d'un arrêté (Arrêté du 11 mars 1968 portant règlement des archives hospitalières). Tant en ce qui concerne son contenu que son archivage, les pratiques des hôpitaux publics ont précédé la législation et la réglementation, comme en témoigne le versement aux archives départementales de dossiers dès 1950 [4].
De 1970 à nos jours la législation et le réglementation concernant le dossier médical vont devenir de plus en plus précises, nous nous limiterons à en citer les principales évolutions. La loi hospitalière nn° 70-348 du 31 janvier 1970 et le décret nn° 74-230 du 7 mars 1974 définissent les modes et les règles de sa communication, au patient, elle ne peut se faire que par l'intermédiaire d'un médecin. La loi hospitalière nn° °91-748 du 31 Juillet 1991 complétée par le décret du 30 Mars 1992 impose la création d'un dossier médical pour les patients hospitalisés et donne une première définition de son contenu.
La loi du 4 Mars 2002 relative aux droits des patients place le patient au centre de la gestion de la communication de son dossier, ceci a peu d'impact sur le dossier "papier" mais induit des contraintes supplémentaires voire des limites à son informatisation. Le décret nn° °2002-637 du 29 avril 2002 étend l'obligation de disposer d'un dossier médical aux patients pris en charge en consultation externe et adapte le contenu en fonction des nouvelles dispositions de la loi :
"Un dossier médical est constitué pour chaque patient hospitalisé dans un établissement de santé public ou privé. Ce dossier contient au moins les éléments suivants, ainsi classés :

1. Les informations formalisées recueillies lors des consultations externes dispensées dans l'établissement, lors de l'accueil au service des urgences ou au moment de l'admission et au cours du séjour hospitalier, et notamment :

   (a)La lettre du médecin qui est à l'origine de la consultation ou de l'admission ;
   (b) Les motifs d'hospitalisation ;

4. http ://www.archinoe.net/cache/sante_affaires_sociales_hopitaux_hospices_poitiers_1789-2003.pdf

(c) La recherche d'antécédents et de facteurs de risques ;

(d) Les conclusions de l'évaluation clinique initiale ;

(e) Le type de prise en charge prévu et les prescriptions effectuées à l'entrée ;

(f) La nature des soins dispensés et les prescriptions établies lors de la consultation externe ou du passage aux urgences ;

(g) Les informations relatives à la prise en charge en cours d'hospitalisation : état clinique, soins reçus, examens para-cliniques, notamment d'imagerie ;

(h) Les informations sur la démarche médicale, adoptée dans les conditions prévues à l'article L. 1111-4 ;

(i) Le dossier d'anesthésie ;

(j) Le compte rendu opératoire ou d'accouchement ;

(k) Le consentement écrit du patient pour les situations où ce consentement est requis sous cette forme par voie légale ou réglementaire ;

(l) La mention des actes transfusionnels pratiqués sur le patient et, le cas échéant, copie de la fiche d'incident transfusionnel mentionnée au deuxième alinéa de l'article R. 666-12-24 ;

(m) Les éléments relatifs à la prescription médicale, à son exécution et aux examens complémentaires ;

(n) Le dossier de soins infirmiers ou, à défaut, les informations relatives aux soins infirmiers ;

(o) Les informations relatives aux soins dispensés par les autres professionnels de santé ;

(p) Les correspondances échangées entre professionnels de santé.

2. Les informations formalisées établies à la fin du séjour :
" Elles comportent notamment :

(a) Le compte rendu d'hospitalisation et la lettre rédigée à l'occasion de la sortie ;

(b) La prescription de sortie et les doubles d'ordonnance de sortie ;

(c) Les modalités de sortie (domicile, autres structures) ;

(d) La fiche de liaison infirmière.

3. Informations mentionnant qu'elles ont été recueillies auprès de tiers n'intervenant pas dans la prise en charge thérapeutique ou concernant de tels tiers.

Ce contenu réglementaire du dossier médial est celui actuellement en vigueur, c'est le minimum obligatoire qui doit être adapté selon le type de prise en charge et selon la

discipline médicale dont relève le patient. D'autres informations peuvent parfaitement intégrer le dossier dès lors qu'elles sont utiles à la prise en charge du patient et qu'il y a nécessité de les partager avec les membres de l'équipe de soins [CNEH, 2011].

1.1.2.1.1.1 Dossier hospitalier en pratique

Il contient de fait tous les éléments d'informations qui sont ou ont été utilisés lors des prises en charge dans l'établissement concerné, il peut s'agir de documents, d' images (radiographies, photographies, ...) ou d'enregistrements . Les explorations et examens à l'origine de ces informations ont pu être effectués lors des hospitalisations, en consultation externe ou lors de prises en charge extra hospitalières (il s'agit alors de copies). Son contenu initialement se limitait à quelques feuillets, parfois associés à 1 ou 2 clichés radiologiques, et à quelques tracés, actuellement il peut correspondre à des dizaines de pages, de nombreux clichés et tracés, et des supports numériques (CD, DVD). Son volume est bien entendu très variable, fonction des pathologies, des modes et du nombre de prises en charge, et de l'existence ou non dans l'établissement d'un DPI. La gestion des dossiers papiers est très variable, fonction des établissements voire des services hospitaliers. On peut distinguer 2 modes de gestion principaux :

- le dossier papier unique pour l'établissement ;
- le dossier papier de service.

Le dossier unique pour l'établissement est la solution "*préconisée*" par la HAS pour la certification des établissements. La difficulté d'assurer la disponibilité des dossiers dans les centres hospitaliers composés de plusieurs établissements, les modifications organisationnelles et des pratiques qu'induisent sa mise en place ont été des freins pour beaucoup d'hôpitaux, qui restent aux dossiers de service. Ces dossiers posent le problème de la partition de l'information concernant le patient dans plusieurs dossiers et de leur coût de gestion et d'archivage.

1.1.2.1.2 Dossier médical de la médecine extrahospitalièere

Alors qu'il n'était jusque là soumis à aucune réglementation, ni de la part du législateur, ni de la part de l' autorité ordinale, un cadre réglementaire apparaît en 1995 au travers de l'article 45 du Code de Déontologie : *"Indépendamment du dossier de suivi médical prévu par la loi, le médecin doit tenir pour chaque patient une fiche d'observation qui lui est personnelle ; cette fiche est confidentielle et comporte les éléments actualisés, nécessaires aux décisions diagnostiques et thérapeutiques. Dans tous les cas, ces documents sont conservés sous la responsabilité du médecin. Tout médecin doit, à la demande du patient ou avec son consentement, transmettre aux médecins qui participent à sa prise*

en charge ou à ceux qu'il entend consulter, les informations et documents utiles à la continuité des soins". [Moutel, 2009]. Aussi novateur soit-il, cet article n'en reste pas moins minimaliste car il n'apporte pas de définition concrète du dossier médical "de ville", ni sur le fond ni sur la forme, c'est à dire aucune information ni sur le contenu, ni sur le contenant.

1.1.3 Limites du dossier médical papier : Difficultés et Attentes

La répartition des informations de santé d'un patient dans de multiples dossiers au sein des services hospitaliers, des établissements, des cabinets médicaux et paramédicaux (voir Figure 1.2), a de nombreux inconvénients dont les plus importants sont :

- l'impossibilité d'avoir une vision complète du passé médical des patients ;
- la réplication des éléments de dossier ;
- la difficulté d'accès à l'information dans le cadre de l'urgence et pour les professionnels nouvellement impliqués dans la prise en charge.

Une autre difficulté est l'accès à l'information dans des dossiers volumineux, que les professionnels ont tenté d'améliorer par le classement des documents par type ou par spécialité. Ces méthodes ont montré leur limite avec l'augmentation du volume des dossiers.
L'informatisation du dossier patient sous forme d'un dossier partagé au mieux unique pour un système de santé, peut améliorer la disponibilité de l'information en préservant son unicité. L'accès à la bonne information au bon moment nécessite en outre des fonctions de visualisation des informations adaptées et des fonction de recherche d'information.

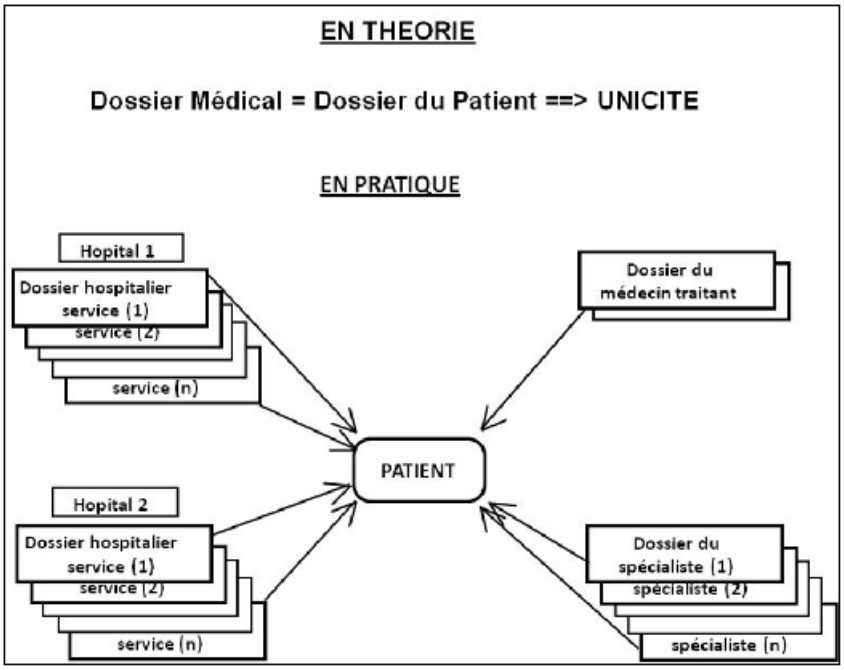

FIGURE 1.2 – Répartition de l'information de santé d'un patient

1.2 L'informatisation du dossier de santé

Les progrès des technologies de l'information ont permis, depuis une trentaine d'année, le développement progressif de dossiers patients informatisés en complément ou en remplacement des dossiers papiers. Les applications en permettant la gestion s'intègrent dans les SIC , partie des SIH et dans le futur dans les SIS .
L'informatisation du DPI offre de multiples opportunités d'amélioration de la qualité des soins : facilité d'accès aux données, communication des dossiers, accès à des banques de connaissances ou à des systèmes d'aide à la décision en lien avec le dossier... Ceci permet également de faciliter l'exploitation des données à visée épidémiologique ou comptable, ce qui est souvent pour les hôpitaux la raison de faire.
Nous scinderons cette section en 3 parties :

- Qu'est-ce que le DPI ?
- La typologie et l'organisation des données dans le DPI ;
- L'état de lieux de l'informatisation du dossier patient.

1.2.1 Qu'est-ce que le DPI ?

Les termes utilisés pour parler de dossier informatisé dans le domaine de la santé sont multiples et recouvrent des concepts différents. Á coté d'*Electronic Health Record(EHR)*, et de sa traduction française *Dossier de Santé Electronique(DSE)*, d'autres termes recouvrant tout ou une partie du même concept sont utilisés : *Electronic Medical Record(EMR)*, *Dossier Médical Electronique(DME)*, *Dossier Médical Partagé(DMP)*,... [Fraser et al., 2005; Häyrinen et al., 2008]. Des institutions, des organisations et de nombreux auteurs ont essayé de préciser les termes et leurs définitions :

L'Organisation Mondiale de la Santé (OMS) [5] définit l'*EHR* comme : "un dossier de santé contenant toutes les informations de santé concernant un individu ; enregistrées et accessibles électroniquement par les soignants de santé tout au long de la vie de la personne ; et couvrant outre les hospitalisations, toutes les situations amenant à lui prodiguer des soins".

La société HIMSS (Health Information Management System) [6] a défini l'*EHR* comme : "un dossier électronique contenant les informations de santé du patient, produit durant les épisodes de prestation de soins. Sont comprises dans ces informations, les données démographiques, les notes, les problèmes, les médicaments, les signes vitaux, les antécédents, les vaccinations, les résultats biologiques et les comptes-rendus radiologiques,...".

5. http ://whqlibdoc.who.int/wpro/2006/9290612177_eng.pdf
6. http ://www.himss.org

L'Organisation Internationale de Normalisation (ISO) [ISO, 2005] définit l'EHR comme : un outil de stockage des données des patients sous forme numérique, dont le stockage et la communication sont sécurisés, accessibles par les utilisateurs autorisés. Il contient des informations rétrospectives (une vue historique de l'état de santé et des soins effectués), courantes (une vue de l'état de santé et des soins en cours), et prospectives (une vue future des soins planifiés) dans le but principal de permettre la continuité, l'efficience et la qualité des soins.

[Degoulet and Fieschi, 1991] définissent le DPI comme suit : Le dossier du patient ne se résume pas à l'observation écrite du médecin (le dossier médical proprement dit) ou aux notes de l'infirmière (le dossier infirmier). Il englobe tout ce qui peut être mémorisé chez un malade, des données démographiques aux enregistrements électro-physiologiques ou aux images les plus sophistiquées. Compte tenu de ce rôle, le dossier du malade est et restera longtemps l'outil principal de centralisation et de coordination de l'activité médicale.

Selon [Lehmann and Meyer zu Bexten, 2002], l'*EHR* est une collection de documents concernant le patient, provenant de différentes sources et créée à différents moments et qui permet, par exemple, de documenter quand, avec qui, pourquoi, quel médicament a été donné, par qui, avec quel résultat et avec quelle argumentation.

En l'absence de définition universellement reconnue, nous utiliserons dans ce travail le terme de *Dossier Patient Informatisé* ayant pour acronyme `DPI`. Les DPI peuvent être différents selon le contexte et la manière dont ils sont mis en oeuvre d'un pays à l'autre, en fonction de l'établissement de santé, voire du mode d'exercice (public ou libéral) ou de la spécialité médicale [Haux, 2006].

1.2.2 DPI : Typologie et organisation des données

Le contenu du DPI en matière d'exhaustivité de l'information, du type de données et de leur organisation conditionne les utilisations des dossiers informatisés : prises en compte des éléments de dossiers, visualisation, recherche d'information, exploitation des données, utilisation des données par des systèmes d'aide à la décision ou l'accès contextuel à la connaissance. Dans ce chapitre nous présenterons les types de données du DPI et leurs principaux modes d'organisation.

1.2.2.1 Typologie des données

Traditionnellement, les professionnels de santé privilégient la culture de l'écrit, ceci explique comme nous l'avons vu dans l'historique du dossier médical, que le dossier médical soit le plus souvent décrit comme un ensemble de documents. Dans un contexte où coexistent le papier et le numérique, les informations peuvent être contenues dans des documents, mais aussi prises en compte sous forme de données structurées ou codées. Généralement, deux types de formulaires utilisés dans la pratique médicale :

- ***Texte libre*** : Les comptes-rendus d'hospitalisation ou d'actes, les notes et commentaires sont en texte libre. Certains de ces documents ont une structuration a minima, plus standardisés lorsqu'ils sont réalisés à partir de trames ou de formulaires, ils peuvent dans certains cas être semi structurés, ils ne répondent que rarement aux normes CDA HL7 ;
- ***Données structurées*** : Certaines données sont enregistrées sous forme de données structurées type attribut-valeur le plus souvent dans le cadre de dossiers de spécialité, les résultats biologiques sont de ce type ainsi que certaines données démographiques. Les traitements médicamenteux lorsqu'ils sont pris en compte sont sous cette forme ;
- ***Données codées*** : Les actes réalisés et les diagnostics des séjours sont codés, respectivement par la CCAM et la CIM 10, plus dans un but médico économique que de soin. L'utilisation du codage à visée indexation est rare, peu d'applications de gestion du DPI offrent les possibilités de le faire et intègrent les outils ou fonctions nécessaires.

Les DPI hospitaliers peuvent le plus souvent prendre en compte la totalité des éléments du dossier médical décrit dans le décret d'avril 2004, à l'exception des traitements médicamenteux en l'absence de CPOE et des notes quotidiennes. Les DPI extra hospitaliers à l'inverse se limitent souvent à une suite de notes et aux ordonnances.

1.2.2.2 Les différents types d'organisation

Selon [Silberzahn, 1997], deux niveaux d'organisation peuvent exister dans les applications du DPI :

- La gestion du contenu avec : soit une approche orientée données, soit une approche orientée documents ;
- L'accès : à ce contenu par une approche orientée vue qui facilite la génération de vues pouvant être adaptées en fonction des besoins des utilisateurs.

1.2.2.2.1 Approches orientées donnée et document

Traditionnellement, l'ensemble des applications de gestion des DPI est orienté selon ces deux approches. L'organisation et la structuration des données du DPI ont pour objectif de faciliter le travail des soignants, et de faciliter l'exploitation des données saisies et enregistrées, en particulier pour le pilotage médico-économique et la recherche.
Dans les **applications orientées documents** quelque soit leur niveau de structuration (documents structurés type CDA [Dolin et al., 2006], des documents semi structurés [Flory et al., 2000], des documents non structurés [Pereira et al., 2009], le texte est l'élément clé du dossier. Dans la majorité des cas, l'information est saisie en langage naturel sous forme libre ou semi structurée. Dans cette approche orientée documents, le DPI est vu comme une collection de documents.
Dans les **applications orientées données**, le dossier est un ensemble de données, l'histoire médicale du patient peut être décrite complètement par ces données et leurs relations. Ces applications permettent de saisir des données par l'intermédiaire d'interfaces de type formulaires structurées et figés : structurés pour faciliter le traitement des données par le système *a posteriori* et figés pour ne donner que peu ou pas de liberté de saisie. Les champs de saisies peuvent être associés à des contraintes pour assurer la cohérence et la validité de ces saisies.
Outre ces fonctions de saisie, ces applications utilisent des vues pour présenter les données et les documents.

1.2.2.2.2 Approche orientée vue

L'évolution de la médecine a conduit à une augmentation de la quantité et la complexité des données médicales, qui a pu faire écrire que le DPI est devenu un `"write once read never (WORN)"` [Powsner et al., 1994].
Donner accès aux données DPI par des fonctions faciles à utiliser et totalement paramétrables et adaptées aux besoins du médecin peut rendre faux cet adage. Plusieurs approches ont été implémentées pour une présentation structurée permettant de clairement identifier, visualiser l'information pertinente dans l'historique du patient et ainsi aider le soignant dans son travail.

- La **structuration suivant la source** *("oriented-source")* qui consiste à regrouper les informations selon leur provenance (établissements, services, spécialités) ;
- La **structuration en fonction du temps** *("oriented-time")*, la plus simple et la plus utilisée, dans laquelle les informations sont représentées chronologiquement sous forme d'évènements. Cette vue est très intéressante lorsqu'on manipule des dossiers de faibles volumes (moins de vingt prises en charge) mais la recherche d'information devient difficile au delà de cette limite ;
- La **structuration par problème** *"Problem Oriented Medical (POM)"* concept

introduit par [Weed, 1971], dont l'originalité du concept qui a pour objectif de permettre la visualisation du DPI en regroupant les éléments par problème de santé. Un *"problème"* est une notion assez large qui peut aller d'une plainte de patient à un diagnostic en passant par des raisons d'hospitalisation ou de consultation [Baud et al., 1998]. Dans cette approche, les vues se fondent sur un **"Gold Standard"** des problèmes médicaux [Rector, 1999](une liste de problèmes rangés hiérarchiquement). Pour obtenir ce résultat, les données sont classées suivant une méthode dont l'acronyme est S.O.A.P :

+ des éléments subjectifs (S), par exemple les informations apportées par le patient ;

+ des éléments objectifs (O) issus de l'observation du médecin (examen clinique et des examens complémentaires) ;

+ l'appréciation (A) (assessment) du problème ;

+ le plan (P) de prise en charge du problème.

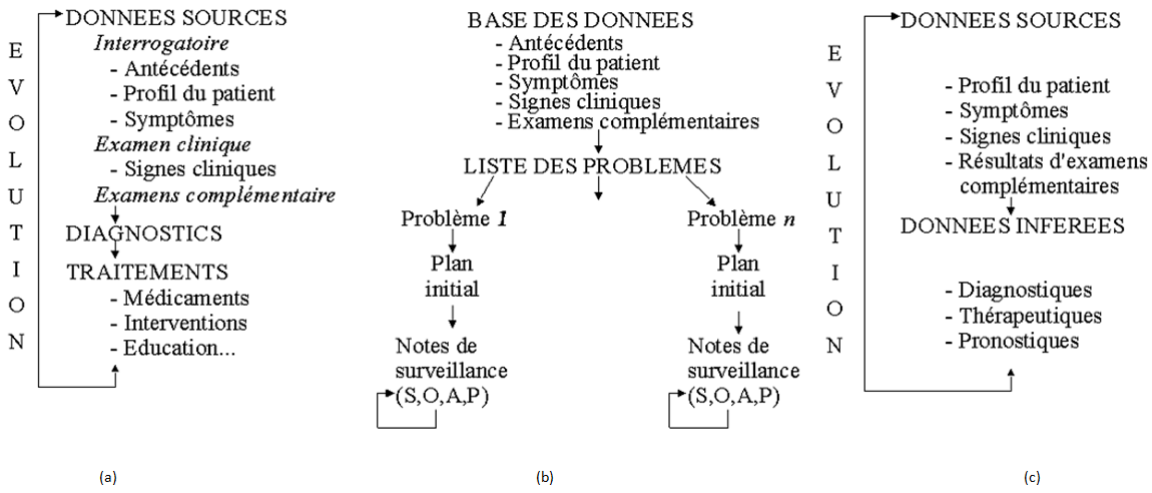

FIGURE 1.3 – (a) Dossier médical orienté SOURCE - (b) Dossier médical orienté PROBLÈME - (c) Dossier médical sémantique et temporel dans [Degoulet and Fieschi, 1991]

Les vues orientées-problèmes peuvent être généralisées en une structuration par concept. Les travaux implémentant la vue *"oriented-concept"*, utilisent majoritairement des ontologies comme base de connaissances ([Bayegan et al., 2002; Zeng et al., 2002; Zillner et al., 2008; Wong et al., 2010]). [Arguello et al., 2009] utilisent les techniques du Web Sémantique (WS) et le modèle HL7 Clinical Document Architecture (HL7 CDA) pour fournir des interfaces qui permettent aux cliniciens de visualiser les procédures médicales effectuées et comment les résultats cliniques ont évolué au fil du temps.

Dans la même optique, dans une approche hybride, [Zeng and Cimino, 2000] fusionnent les approches précédentes pour montrer leur complémentarité, l'approche "oriented-scenario" d' [Yousefi et al., 2009] permet d'extraire des informations spécifiques à une maladie selon des scénarios cliniques.

1.2.2.3 Le DPI au sein des SIH

[Collen, 1991], décrit l'évolution des pratiques des PDS par l'adoption d'outils informatiques les aidant dans la collecte, la saisie, le stockage et la recherche des données cliniques. Cette évolution a débuté aux États-Unis en 1970 par l'informatisation des SIH. En France, la mise en oeuvre de cette informatisation des SIH [Degoulet et al., 2003; Fieschi, 2003] s'est déroulée en plusieurs étapes, d'abord limitée à la gestion administrative de l'hôpital, puis étendue à l'aspect médico-technique et aux soins par le développement des DPI. Ainsi faire l'état des lieux du DPI hospitalier est indissociable de celui des SIH. L'état des lieux des DPI extra hospitaliers sera fait indépendamment.

[Degoulet and Fieschi, 1991] définissent le SIH comme *"un environnement logiciel et matériel destiné à faciliter la gestion de l'ensemble des informations médicales et administratives de l'hôpital"*. Cette définition rend le DPI partie intégrante des SIH. Les SIH, comme tout autre SI, peuvent être décrits selon 3 dimensions [Reix, 2004] :

- La dimension **informationnelle** concerne la circulation des données et des connaissances au sein des organisations et entre les organisations ;
- La dimension **organisationnelle** concerne les processus de travail (réalisation des tâches, coordination "verticale" ou "transversale" des tâches, ...) ;
- La dimension **informationnelle** concerne les technologies utilisées (stockage, traitement, ...).

Les dimensions informationnelles et technologiques impactent fortement le DPI, la dimension organisationnelle impacte plus fortement les autres composants du SIC (applications de prescription, gestion des processus de soin, gestion des rendez-vous en particulier). [Fieschi, 2003] distingue deux types de besoins informationnels correspondants respectivement au mode organisationnel à certains systèmes : *On Line Transaction Processing* (OLTP) pour l'utilisation des données du DPI durant le processus de soins et *On Line Analytical Processing* (OLAP) pour la ré-utilisation des données DPI à des différentes fins (voir Figure 1.4).
D'un point de vue informationnel, le DPI permet le partage des informations utiles non seulement à la prise en charge du patient, mais aussi à la gestion économique et de la santé publique.

D'un point de vue technologique, les DPI sont très dépendants des technologies utilisées par les autres applications du SIH. Des avancées existent dans le domaine des SIH sur la représentation des données et des connaissances permettant de faciliter le traitement de l'information et l'interopérabilité des systèmes : la normalisation des communications inter applicatives, les applications de communication, la standardisation des données et l'utilisation d'entrepôts de données. Un des problèmes de la mise en place d'application de gestion de DPI est leur intégration au sein des SIH.

Néanmoins la dimension organisationnelle intervient sur le mode et le périmètre de partage des données.

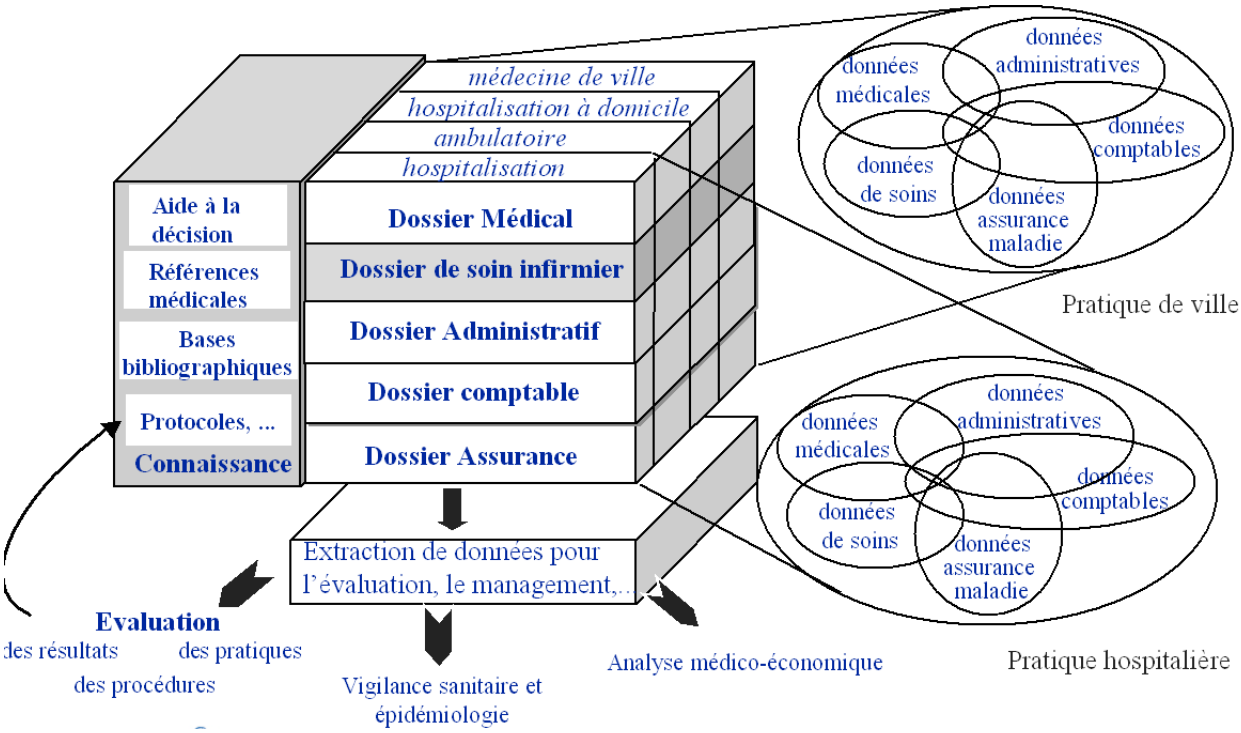

FIGURE 1.4 – Dimension informationnelle des SIH [Fieschi, 2003]

1.2.2.4 Mode et périmètre de partage des données

Les premières expériences de mise en place de DPI au sein des SIH, se sont faites sur un mode dit "*vertical*" par juxtaposition d'applications correspondant à certains services de spécialité de l'établissement. Ces applications sont souvent reconnues sous le terme de *dossiers de spécialité*.

1.2.2.4.1 Dossier de spécialité (DS)

Les services hospitaliers ont à leur disposition leur propre outil informatisé permettant de gérer des dossier "*métier*" adaptés à leurs pratiques. Le dossier de spécialité est un dossier qui contient toutes les informations médicales concernant un patient, relatives à la spécialité mais aussi des informations de santé plus générales. Les données qu'il contient sont gérées en grande partie au sein de formulaire et de type *attribut-valeur*. Une des évolutions a consisté à définir au sein de ces dossiers de spécialité, un corpus de données partageables avec les autres services des établissements, aboutissant au dossier minimum commun.

1.2.2.4.2 Dossier minimum commun (DMC)

Selon [Lukacs and Lang, 1989], le DMC comprend outre l'identification du patient, des informations cliniques de synthèse consensuellement définies au sein d'un établissement, ces informations sont souvent limitées aux séjours dans l'établissement, aux pathologies diagnostiquées et aux données d'importance vitale.
Outre les difficultés de la définition des éléments partageables [Zweigenbaum, 1999; Fieschi, 2003], le DMC a rapidement montré ses limites. La mise en place de dossiers de spécialité dans tous les services d'un hôpital, n'a été que rarement possible du fait des coûts ou de l'absence sur le marché d'applications adaptées et intégrables au SIH. L'informatisation des services "*non équipés*" s'est faite en recourant à des applications bureautiques. L'existence de données numériques, le plus souvent sous forme de documents, a été l'opportunité d'utiliser l'outil informatique à des fins de communication des éléments de dossiers médicaux entre les services d'un même établissement.

1.2.2.4.3 Dossier médical partagé

[Fieschi, 2003] définit le terme de dossier médical partagé, comme un réservoir commun de données, dont les données sont accessibles dans un espace donné, qui peut être un hôpital, un réseau de soin voire le système de santé d'une nation, il s'agit du DMP (Dossier Médical Personnel)pour le SIS français.

1.2.3 État des lieux de l'informatisation du dossier patient

Dans la plupart des applications de gestion des DPI coexistent, les approches orientée-documents et orientée-données , c'est par exemple le cas du progiciel Cpage Dossier Patient (CDP) du CHU de Rouen qui permet de gérer le DPI en rattachant au patient un ensemble de données, et documents avec 4 niveaux (patient, épisode, séjours, actes) [Massari and Fuss, 2000]. Les applications du marché de type PGI (progiciel de gestion intégré), comprennent des fonctions de gestion du DPI de ce type. Les solutions proposant un *patchwork* applicatif sont plus orientées données, mis à part les éléments partagés qui sont majoritairement sous forme de documents. Ainsi dans ces systèmes, les informations sont principalement contenues dans des documents, sous forme non structurées et non codées.
Ces applications utilisent des vues pour présenter les données et les documents, les vues se limitent dans bien des cas à des vues chronologiques et à quelques vues de gestion. La possibilité de gérer des dossiers orientés problèmes existe dans certaines applications, la lourdeur de la gestion de dossier en a toujours limité l'usage [Fakoff, 1999]. Les vues "*métiers*" facilitant l'accès aux documents ou aux informations dans le cadre du soin, se limitent le plus souvent à des démarches expérimentales.)). Le DPI du CHU de Rouen propose des vues de ce type en utilisant les métatermes issus de la terminologie CISMeF [Massari et al., 2008].
Les DPI hospitaliers peuvent le plus souvent prendre en compte la totalité des éléments du dossier médical décris dans le décret d'avril 2004, à l'exception des traitements médicamenteux en l'absence de CPOE et des notes quotidiennes. Progressivement le DPI partagé au sein d'un établissement s'impose, c'est une des conditions permettant d'inclure les documents produits dans le DMP. Le DMP est géré par des applications orientées documents, ces documents vont progressivement être à la norme IHE CDA. Des travaux sont en cours [7] afin de permettre de disposer de vues facilitant l'accès aux informations.
Les données fiables sur le taux d'hôpitaux français ayant mis en place un DPI sont rares ou anciennes. L'interprétation des chiffres est de plus difficile, la différentiation entre dossiers de spécialités et dossiers "d'établissement" est rarement faite, le nombre d'unités ou de services l'utilisant est rarement mentionné. La plupart des hôpitaux ont mis en place ou sont en cours de mise en place d'un DPI, souvent dans des cadres plus vastes, incluant en particulier l'informatisation de la prescription. Le DPI existe depuis près de 20 ans dans certains hôpitaux, les dossiers de certains patients deviennent très volumineux, ce qui rend l'utilisation des informations qu'ils contiennent difficile. Dans notre expérience, 33.46% des dossiers de CDP contiennent plus de 20 séjours.
L'informatisation des dossiers extra hospitaliers a évolué parallèlement, au cours des an-

7. Le projet LEcture Rapide en Urgence du Dossier Informatisé du patient (LERUDI), URL : http ://esante.gouv.fr/dossiers/le-projet-lerudi-fiche-signaletique

nées 1990, nous avons assisté en ville à une informatisation quasi générale des cabinets des médecins généralistes [Marc et al., 2009]. Les motifs en sont simples et tiennent en grande partie à la mise en place de la "Feuille de Soins Electronique (FSE) par l'ordonnance Juppé 96-345 du20 avril. Une enquête de 2008 sur les services électroniques dans le secteur des soins de santé (eHealth) [8], a estimé à 82.8% le nombre de médecins français informatisés mais ce taux est en progression constante en fonction des avancées technologiques opportunes (pour la France). La gestion informatique reste, en ville, individuelle et peu communicante. Les DPI extra hospitaliers à l'inverse se limitent souvent à une suite de notes et aux ordonnances.

8. http ://ec.europa.eu/luxembourg/news/frontpage_news/179_fr.htm

Synthèse

Dans ce chapitre, nous avons présenté l'historique du dossier médical et ses évolutions en médecine hospitalière et médecine extra hospitalière. La multiplication des documents et la prise en charge partagée par plusieurs professionnels parallèlement aux progrès des TIC, ont rendu inéluctable le développement des DPI. Ces DPI permettent le partage des données [et des documents] à travers la mise en place de dossiers médicaux partagés et offrent des vues plus ou moins adaptées à l'action médicale.
Étant donné le grand volume d'informations à manipuler dans les dossiers patients, les DPI, outre les fonctions de communication, doivent permettre la recherche d'information et l'exploitation des données.
Ceci nécessite le développement d'outils de RI proches des outils existants pour la RI documentaire (cf. chapitre 3), ces outils ne peuvent fonctionner que dans des applications ayant un modèle de données adapté à cette RI et gérant l'information sous une forme *structurée* et *standardisée* (cf. chapitre 2) *a priori* ou *a posteriori* à l'aide d'outils TAL .

CHAPITRE 2

Modélisation du DPI

Introduction

Les modèles utilisés pour stocker et organiser les informations médicales contenues dans le dossier patient, étaient adaptés aux contraintes fonctionnelles, organisationnelles et techniques des SIC parmi lesquelles : la rapidité et facilité d'utilisation du dossier patient par les utilisateurs, l'aide à la pratique médicale, les recherches cliniques, épidémiologiques, et l'intégration d'outils d'aide à la décision [Gouveia-Oliveira and Lopes, 1993; Degoulet and Fagon, 2004]
Une grande partie des informations médicales de l'histoire du patient sont sous forme texte dans les courriers, les comptes-rendus d'hospitalisation et d'acte et les notes personnelles du médecin,.... Pour qu'une information soit utilisable et accessible, elle doit être indexée à l'aide de vocabulaires de référence [Rector et al., 1991] grâce à des outils TAL fondés sur les principales terminologies et ontologies de santé.
Le dossier de santé même informatisé ne contient pas de manière explicite l'ensemble des informations, une grande partie de l'information critique est implicite et le lien de causalité (s'il existe) n'y est pas directement mentionné [Johnson et al., 1991]. Par exemple, quand un médicament doit être arrêté en raison de ses effets secondaires (par exemple : anémie due à la chimiothérapie), le lien de causalité entre effets secondaires et arrêt du médicament est rarement mentionné, et la nature de l'effet secondaire, lui-même doit être déduit des résultats d'analyses biologiques car il n'est pas explicitement statué comme "anémie".
Pour cela, il faut reconstruire l'histoire *"chronologique et sémantique"* de santé du patient à partir de l'ensemble des informations en identifiant les différents concepts et leurs

relations.
Disposer de données structurées et indexées est une condition préalable au développement d'outils de RI mais, ce n'est pas suffisant. Ces données devront être enregistrées dans un modèle d'information [de données] adapté à la RI dans notre cas, indépendamment du modèle d'information du système de gestion des dossiers patient. Ce point est central car l'objectif de notre modèle est de pouvoir s'adapter à n'importe quel DPI.
L'état de l'art de la problématique de modélisation est résumé dans la figure 2.1.

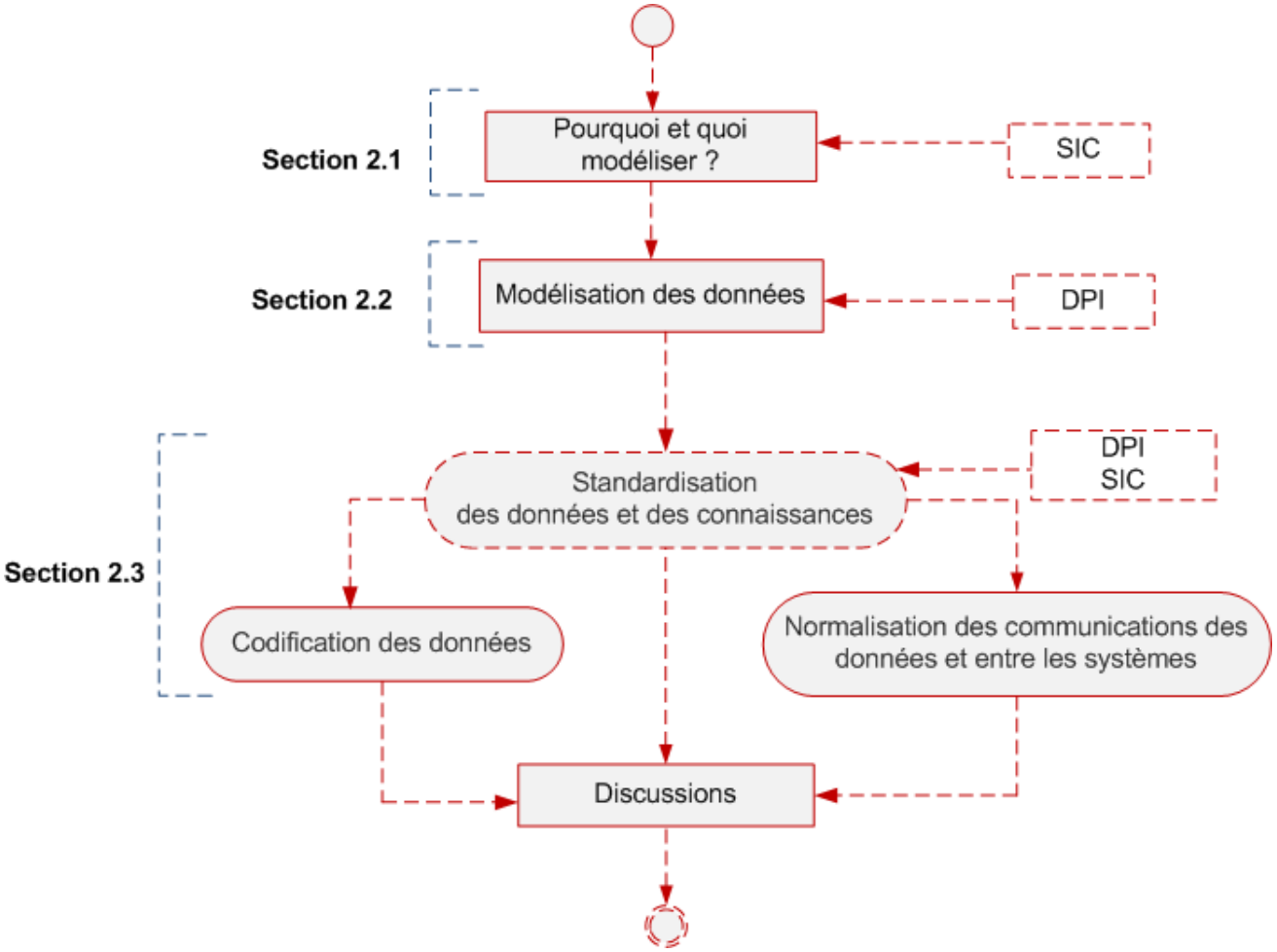

FIGURE 2.1 – Schéma synoptique des différentes modélisations

2.1 Pourquoi et quoi modéliser ?

Un modèle peut être défini comme une représentation abstraite de concepts, renonçant au détail, qui a pour but d'isoler et de préciser un ensemble de propriétés fondamentales [Degoulet and Fieschi, 1991]. Par conséquent, la conception de tout système informatique dépend des concepts qu'il gère (et/ou manipule) ainsi que de leurs relations, et des traitements effectués qui vont permettre de définir le modèle le plus adapté [Essin and Lincoln, 1994]. Le moyen le plus simple de représenter et de transmettre des informations et des connaissances est le langage naturel. Des modèles spécifiques ont été développés pour des besoins propres, comme par exemple les modèles statistiques ou probabilistes, les modèles linguistiques, les modèles graphiques pour les connaissances géographiques. Quant au modèle informatique, il a permis de dissocier les données et leur signification [Degoulet and Fieschi, 1991]. Par exemple, l'expression en langage naturel "Pierre, âgé de 40ans, est traité par l'aspirine" est interprétée comme "le traitement médicamenteux de Pierre (qui est une personne de sexe Masculin) est l'Aspirine". Un programme informatique stocke la donnée "Aspirine", dans le traitement d'un patient nommé "Pierre".
L'enregistrement de ces informations et leur traitement nécessitent de définir préalablement des modèles de représentation des données (*data modelling*) et de leurs significations (*concept modelling*) [Rector et al., 2001].
Plusieurs modèles ont été conçus pour décrire les SIC (*a fortiori* les DPI). [Rector et al., 2001] les catégorisent en 3 types de modèles (voir Figure 2.2) :

- les **"modèles d'information [de données]"** pour modéliser les informations spécifiques au patient *(1)* ;
- les **"modèles terminologiques [ou ontologiques]"** pour modéliser les éléments décrivant le sens de ces informations *(2)* ;
- les **"modèles d'inférences"** pour modéliser les connaissances nécessaires (sens règles) pour en déduire des conclusions, des décisions, et des actions à suivre sur ces informations *(3)*.

Pour couvrir et représenter la complexité du dossier patient, [Goble et al., 1994] décrit deux autres modèles dans une approche orientée *"activité/prestation"* :

- les **"modèles de processus"** pour représenter les processus de soins (ensemble des activités) ;
- les **"modèles de l'utilisateur"** pour représenter les acteurs et leurs profils dans leurs interactions avec chacun d'eux et avec les autres modèles.

Les raisons qui rendent nécessaires ces modèles sont nombreuses, au delà de *l'historique d'un modèle* (chaque modèle possède une histoire par exemple, le modèle traditionnel

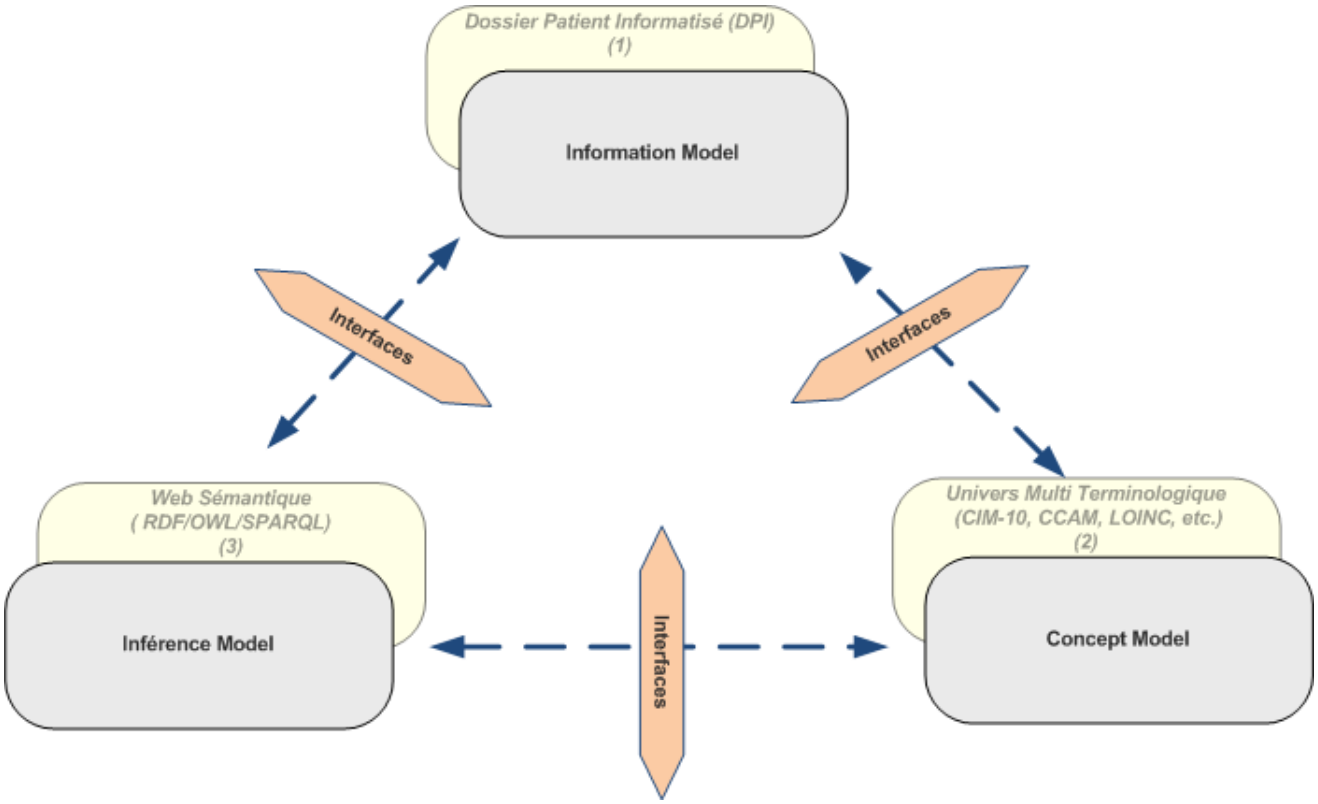

FIGURE 2.2 – Instanciation du modèle de [Rector et al., 2001] dans notre contexte

'Entité-Relation' pour les bases de données relationnelles) et son *adaptation organisationnelle* (par exemple : nécessité d'échange et de partage des données entre différents acteurs et systèmes), [Rector et al., 2001] définissent deux raisons importantes :

- **"Technique"** pour définir le mode et format de stockage ;
- **"Facteurs humains et ergonomiques"** pour prendre en compte l'utilisateur final (comme par exemple au sein d'EVALAB à Lille [1]).

Notre travail se focalise sur la problémantique de la modélisation du DPI. Néanmoins, nous présentons les contraintes inhérentes à cette modélisation à savoir la standardisation des données et des connaissances pour un modèle interopérable d'un point de vue *technique* et *sémantique*.

1. EVALAB, URL : http ://evalab.univ-lille2.fr/fr/presentation-de-l-evalab.html

2.2 Modélisation du DPI

Les modèles historiquement associés à la conception linéaire du stockage des données médicales (celles du dossier médical `papier`) ont montré leur limite [Johnson, 1996]. L'information médicale recueillie, quelque soit son format (texte ou code) durant le processus de soins est difficile à manipuler.
Pour [Essin and Lincoln, 1994], le modèle d'information est nécessaire pour permettre d'adapter la représentation des informations en une représentation informatique compréhensible par la machine qui contiendra ces données. Ces-dits modèles, selon [Rector et al., 1991], vont permettre de représenter explicitement l'information afin de la traiter (extraire, rechercher, agréger,...).
L'évolution des techniques informatiques a permis d'envisager des modèles (ou structures) plus complexes faisant une nette distinction entre le modèle de représentation interne des données (ou "structure profonde") et les vues (ou "structure de surface") que les utilisateurs peuvent avoir de ces données [Degoulet and Fieschi, 1991]. Ces structures ont permis de stocker, de manière explicite, les contextes de ces informations et les connaissances cliniques qui ne font pas partie des données du DPI [Gouveia-Oliveira and Lopes, 1993].
Concevoir un modèle d'information revient à concevoir un modèle sémantique de données, qu'il faut bien distinguer du modèle d'implémentation de ce dernier. Les modèles sémantiques de dossiers médicaux, qui trouvent leur racine dans les travaux en Intelligence Artificielle (IA) et en Analyse du Langage Naturel, visent à expliciter les relations entre les différentes données médicales [Hull and King, 1987]. Il existe différents modèles sémantiques, parmi les plus répandus : le modèle relationnel dont la sémantique est représenté par des contraintes d'intégrité, le modèle orienté objet pour représenter les propriétés structurales et comportementales des objets, les modèles sémantiques utilisant des langages standards comme XML, RDF ou des langages formels comme les logiques de description [Brachman and Schmolze, 1985] ou les graphes conceptuels [Sowa, 1983], ...
La conception d'un modèle sémantique de données dans le cadre de la RI doit prendre en compte deux éléments importants [Johnson, 1996] :

- la nécessaire rapidité de la recherche d'information dans un dossier patient (trouver l'information en un temps compatible avec les contraintes de l'exercice médical). Autrement dit la `scalabilité` (passage à l'échelle) de la RI dépend du contexte : un soignant attendra quelques secondes pour une RI dans un contexte de soins ; il acceptera d'attendre plusieurs minutes, s'il s'agit d'une RI sur une base de dossiers médicaux, dans un contexte de recherche clinique.
- son adaptation à intégrer de nouveaux types de données (les processus de soins évoluent constamment).

Dans les paragraphes suivants nous présenterons les modèles utilisés pour l'informatisation du DPI en individualisant ceux plus spécifiquement développés pour la RI.

2.2.1 Panorama des approches de modélisation

Dans le contexte de la modélisation du dossier médical, les travaux les plus souvent menés ont pour objectif de représenter la sémantique de l'information contenue dans le DPI afin de pouvoir : rechercher, extraire, visualiser une information médicale spécifique à un patient ou d'agréger des informations médicales d'une population de patients à des fins d'analyses.
Les travaux sont nombreux mais aucun modèle d'information standard communément reconnu par le domaine n'existe à ce jour [Campbell et al., 1994; Smith and Ceusters, 2006], néanmoins un modèle multidimensionnel, le modèle I2B2 [Murphy et al., 2006] émerge et est considéré comme un standard de facto (cf. paragraphe 4.2.2.1.4 du chapitre 4).
Les approches de modélisation du DPI sont diverses, elles prennent en particulier en compte : *le contexte, le temps, le processus de soins et l'utilisateur (ou acteur).*
Les soins médicaux étant prodigués par plusieurs acteurs intervenant dans la prise en charge du patient, il n'est plus utile de se poser des questions sur la nécessité d'utiliser un DPI mais sur la manière de prendre en compte les différents points de vue de ses acteurs en utilisant leur langage métier et en prenant en compte aussi ces aspects temporels, contextuels et les processus. Car ceux sont les intéressés [2] comme le soulignent [Dieng-Kuntz et al., 2001].
[Weed, 1971] définit la notion de contexte comme "une pensée, un objet, un événement ou une activité qui est lié à ses contextes ou ses circontances, ..." pour une meilleure compréhension de l'information et pour améliorer le raisonnement sur cette dernière.
[Rector et al., 1993; Campbell et al., 1994; Dolin, 1994] considèrent l'aspect temporel comme la base sur laquelle doit être construite le dossier médical. La modélisation fondée sur la temporalité de l'information *"temporal information"* du DPI est un domaine récent de recherche [Zhou and Hripcsak, 2007].

2. *stakeholder* - ceux sont toutes les personnes ayant un intérêt ou un enjeu qui peuvent être affectés par le système

2.2.2 Des modèles d'information adaptés pour la gestion des données

Les approches, qui prennent en compte le contexte et/ou le temps, utilisent différentes méthodes [formelles et non formelles] pour représenter de manière explicite le contexte et le temps. Les modèles déjà réalisés avec cette approche orientée *"contexte"* sont nombreux avec des méthodes de conception différentes techniquement et conceptuellement :

- Selon une approche *orientée vue*, les modèles ([Gouveia-Oliveira and Lopes, 1993; Barrows Jr and Johnson, 1995]) qui sont adaptés à des contextes spécifiques "oriented-problem context" pour les premiers, "causal context" pour les seconds ;
- Certains utilisent le formalisme des réseaux sémantiques [Rector et al., 1993; Rogers et al., 2006], ces deux modèles seront détaillés, respectivement, dans les paragraphes 4.2.2.1.1 et 4.2.2.1.2 du chapitre 4 ;

- D'autres modèles mettent en oeuvre les théories des graphes conceptuels introduits par [Sowa, 1983] et les logiques de description introduites par [Brachman and Schmolze, 1985] en les implémentant dans différents formats.

Pour modéliser un compte-rendu de radiologie, [Campbell et al., 1994] utilisent les logiques de descriptions dans un graphe conceptuel associé avec un modèle métrique pour représenter la dimension temporelle des données. [Pinon et al., 1997] utilisent le langage XML pour décrire formellement un compte-rendu médical en exploitant la structure logique (les différents blocs) à laquelle une couche sémantique (utilisation d'une ontologie médicale) est associée afin de donner du *sens* au contenu de chaque bloc.
Quant à [Lindemann et al., 2009], ils représentent les données médicales du patient au format de triplets RDF, et enfin [Patel and Cimino, 2007] représentent le contenu du dossier patient sous forme d'assertions (faits) projetées sur la terminologie SNOMED afin de pouvoir mettre en parallèle les données de dossiers patients et des protocoles d'essais cliniques.

- D'autres auteurs ont utilisé des bases de données relationnelles et/ou objets :

[Branson et al., 2008] représentent le dossier médical sous forme d'événements médicaux incluant plusieurs variables cliniques (mesure, annotation, lien source, observation, concept médical, ...) définis par des métadonnées.
[Johnson, 1996] est le premier à justifier le besoin d'un modèle générique et flexible, à la fois adapté à la RI et à l'évolution du domaine. Il propose une technique de modélisation

fondée sur les graphes conceptuels pour arriver à un modèle efficient avec un minimum de tables. Toujours selon lui, cette modélisation pallie aux approches de modélisation traditionnelles des données médicales avec des centaines de tables et un ensemble riche de contraintes, lesquelles sont inefficaces pour les DPIs et pour les requêtes *orientées sur les données du dossier patient.*
Le modèle pragmatique de [Degoulet and Jean, 1989] vise à ajouter d'autres dimensions, en particulier celle du contexte pour améliorer son modèle sémantique et temporel [Degoulet, 1984]. L'information médicale est sous forme de quadruplet de type "objet-attribut-valeur-temps" par exemple <Dupont, pression artérielle systolique,170 mmHg, 10/9/88>.

– Des approches mathématiques sont aussi proposées pour modéliser la complexité temporelle des données médicales [Dolin, 1995; Huff et al., 1995; Hripcsak et al., 2005; Lai et al., 2008]. Les évenements médicaux sont représentés par des intervalles, et les assertions sur les évenements comme des contraintes. Ceci semble nécessaire pour représenter la plupart des assertions temporelles des CR médicaux [Hripcsak et al., 2005].

2.2.3 Des modèles d'information adaptés à la visualisation des données

Des approches de modélisation se positionnent sur l'utilisateur final et sur sa perception de l'environnement informatique, les unes basées sur des approches orientées "vues" [3] du dossier médical, les autres sont fondées sur les pratiques métiers.
[Flory et al., 2006] distinguent, dans son modèle, des objets simples pour représenter les éléments médicaux classiques (visite, ordonnance, biologie, radiographie, etc.) et des objets complexes (représentation d'un processus de soins, par exemple un épisode de soins), afin de proposer une nouvelle interface personalisable de suivi et de gestion du dossier patient.
[Guisiano et al., 1992] qui s'orientent vers un découpage des informations selon des niveaux de généralité déterminant leur intérêt commun ou spécifique pour le médecin dans un besoin de recherche d'information ; et enfin [Huet et al., 2001] propose une métamodèlsation du dossier patient basée sur une approche cognitive afin d'intégrer le langage métier du médecin.

3. Voir sous section 1.2.2.2.2

2.2.4 Des modèles d'information dédiés à la communication

Ces modèles d'information, conçus dans une nouvelle approche conceptuelle semblable à l'approche orientée *"processus"* du monde industriel, associent les processus de soins, les acteurs et le patient dans une architecture globale dans le partage des données du DPI. Ils correspondent à l'évolution de SIH orientés *"patient"* vers des SIH orientés *"activité/prestation"*. Cette orientation répond aux besoins d'interopérabilité des systèmes, autrement dit un moyen d'intégration dans le but de faciliter la communication entre les SIH (*a fortiori* les SIC). Ces modèles sont des `"Reference Model (RM)"` pour avoir une représentation standardisée et internationale du DPI pour les besoins que nous venons de citer.
Dans le cadre du projet de recherche européen Health Telematics Research Program fondé par AIM (Advanced Informatics in Medicine) [4] en 1992, plusieurs travaux ont été entrepris pour mettre en place les fondements d'une architecture formelle pour ce besoin d'échange des données entre les systèmes conformément aux éxigences légales et éthiques [Iakovidis, 1998]. Nous citons ici les principaux standards européens :

- La prénorme ENV 13606 connue sous l'acronyme Electronic Healthcare Record Communications (EHRcom) [Kay and Marley, 1999] ;
- Good European Healthcare Record (GEHR) [Kalra and Lloyd, 1995], une révision de la prénorme ENV13606 vers un modèle normatif complet en adoptant une approche de modélisation multi-niveau fondée sur le concept d'`archétype`(cf. paragraphe **??**) ;
- La prénorme ENV 12967 connue sous l'acronyme HISA (Health Informatics Service Architecture) [Scherrer and Spahni, 1999] qui a fait l'objet d'une normalisation ISO 12967 :2009.

D'autres initiatives ont vu le jour, la norme ISO 18308 :2006 connue sous le signe d'`openEHR`, un modèle standard et open-source pour le développement d'un SIS standard [5] [Beale, 2002] ; La norme ENV 14822 connue sous l'acronyme Reference Information Model(RIM), qui a fait l'objet d'une normalisation ISO 21731 : 2006 [HL7, 2005].
Nous ne détaillons pas certains de ces standards [6] car plusieurs revues de la littérature ont étudié et examiné ces normes pour :

- déterminer les niveaux d'interopéabilité et les fonctions qu'elles fournissent d'un point de vue structure de données, de sécurité de données, ... [Eichelberg et al.,

4. a European Commission research initiative
5. www.openehr.org
6. Seuls les modèles RIM, HISA et openEHR seront détaillés dans le paragraphe 4.2.2.2 du chapitre 4.

2005] ;
- de définir les relations entre ces standards [Schloeffel et al., 2006] ;
- d'évaluer la formalisation de leur modèle [Cuggia et al., 2009] ;
- d'en déduire une démarche conceptuelle pour modéliser les différents domaine des SIS [ANAP, 2010].

Quelques implémentations de ces standards dans la littérature : EHRcom [Ken and Dipak, 2008], GEHR [Bird et al., 2003], HISA [Massari and Fuss, 2000], RIM HL7 [Lyman et al., 2003], openEHR [Kohl et al., 2010].

2.2.5 Des modèles d'information pour l'analyse des données

Face au nouveau paradigme de la médecine translationnelle [Bréchot, 2004], des modèles orientés "Data Warehouse (DW) [7]" sont proposés. Ces modèles [Choquet et al., 2008; Zapletal et al., 2010; Cuggia et al., 2011], sont pour la plupart, implementés dans un schéma physique en étoile et permettent des requêtes multidimensionnelles(requêtes OLAP) dont l'aspect temporel peut constituer une dimension à part.
Le modèle I2B2 [Murphy et al., 2006] permet d'intégrer des données médicales diverses (biologie, génétique, clinique,etc.) afin de pouvoir faire des interrogations multicritères dans le temps. D'autres modèles [8] similaires au modèle I2B2 existent, nous pouvons en citer quelques-uns : OMOP Common Data Model [9], HMORN Virtual DataWarehouse [10], HIMSS Data Model [11].

7. Entrepôt de données
8. A détailler ces 3 modèles, si possible...
9. Observational Medical Outomes Partnership (OMOP), URL : http ://omop.fnih.org/CDMandTerminologies
10. http ://www.hmoresearchnetwork.org/resources/toolkit/HMORN_VDWDataStructures.pdf
11. http ://www.himss.org/content/files/EHRAttributes.pdf

2.2.6 Discussions sur la spécificité des modèles à la RI

La modélisation est un processus fondamental qui conditionne le fonctionnement et les performances d'un SIC. De nombreux travaux ont été consacrés à la modélisation des données médicales.
Des problématiques connexes (*a fortiori* complémentaires) à la RI ont été privilégiées comme :

- l'`exploration du DPI` à travers de vues adaptées à l'utilisateur ou sous forme de synthèse. Ces approches souffrent de certaines limites : un trés gros travail de catégorisation des concepts par problème pour [Zeng et al., 2002]. Bien que cette vue orientée problème améliore la navigation à travers le DPI, elle ne résout par le problème de surplus d'information [Massari et al., 2008] et la navigation implique le fait que l'utilisateur doit connaitre ce qu'il cherche pour formuler sa requête à travers ces interfaces, ce qui n'est pas toujours le cas ;
- la `"[ré]utilisation des données médicales"` (recherche clinique, épidémiologique ou translationnelle). La constitution de cohortes et la recherche épidémiologique peuvent être vues comme des cas particuliers de la RI. Pour [Johnson, 1996], l'accès rapide aux données individuelles du patient est un critère important pour la modélisation.

Les modèles normatifs, proposés dans le cadre de l'interopérabilité des SIC pouvent être utilisés pour structurer les données du DPI mais n'ont pas été conçus pour faciliter la RI. Pour certains auteurs (ou modèles), les problématiques relatives à la RI ont été prises en compte comme le contexte, la temporalité, la sémantique des données médicales afin de les exploiter efficacement. Ces modèles ont été conçus pour des cas d'utilisation spécifiques. Les techniques de RI utilisables sont dépendantes du modèle et bien d'autres critères inhérents à ce dernier (formalisme, méthode d'implémentation, ...). Les motivations liminaires de ces travaux montrent que, dès l'étude et la conception du modèle, les spécificités de la RI ne sont pas pries en compte pour la plupart.
Notre objectif est de concevoir un modèle de données générique et flexible comme pour [Johnson, 1996] adapté à la RI au sein d'un DPI ou dans une base de dossiers médicaux. Nous détaillerons les besoins et les contraintes de conception dans la section 4.2 du chapitre 4.

2.3 Standarisation des données et des connaissances médicales

2.3.1 Codification de l'information médicale

Comme nous l'avons énoncé dans la section 2.2, la modélisation du DPI implique la modélisation des éléments décrivant le sens des données du DPI.
Le langage médical est caractérisé par un vocabulaire extrêmement riche et difficile à manipuler. Les termes utilisés sont souvent trés imprécis et font rarement l'objet de définitions rigoureuses. Dans ce type de langage, il y a plusieurs façons d'exprimer la même chose (synonymies), ainsi que plusieurs interprétations possibles pour des termes (homonymie), en particulier pour les acronymes (IVG :Interruption Volontaire de Grossesse ou Insuffisance Ventriculaire Gauche ?). Il faut donc définir un modèle formel pour traiter l'information médicale par une machine. Différentes approches existent pour modéliser les informations médicales : de la constitution de référentiels de codage pré coordonnés à la mise au point de systèmes formels génératifs permettant de composer à l'infini les concepts élémentaires d'une ontologie [Zweigenbaum, 1999].
[Spackman et al., 1997] classe les systèmes terminologiques en 3 catégories : les terminologies de références (TR) comme celles que nous présenterons dans la sous section 2.3.1.1 pour la RI, le stockage et le traitement des données cliniques ; les terminologies de traitement (TT) pour l'optimisation des TAL et enfin les terminologies d'interfaces (TI) pour la conception d'interfaces de saisie des données mieux adaptées aux besoins informationnels des professionnels [Natarajan et al., 2010] et qui sera développé dans la sous section 2.3.1.2.
Les TR et surtout les TI sont trés utiles pour la RI dans le DPI. Nous détaillons ici les TR utilisées dans le cadre de notre travail.

2.3.1.1 Terminologies de références (TR)

Le contenu et la structure d'une TR dépendent de la fonction pour laquelle cette TR va être utilisée : nécessité de trouver une solution pour une raison légale (obligation de coder les diagnostics pour le PMSI), pour une raison informatique (aide au codage et au traitement), pour représenter des métiers ; d'autres ont été conçues à des fins de recherche et d'exhaustivité médicale et s'apparentent à des ontologies.
Dans le cadre de cette thèse, nous avons utilisé un certain nombre de TR de différents types. La plupart sont traduites en français. Nous définissons dans cette section 5 TR principales pour le développement de nos travaux :

– La **classification CIM-10**(Classification Internationale des Maladies) : patronnée par l'OMS, elle est utilisée à travers le monde pour enregistrer les causes de morbidité et de mortalité, à des fins diverses parmi lesquelles le financement et l'organisation des services. La CIM10 permet le recueil de diagnostics à des fin de santé publique ou d'évaluation de l'activité hospitalière, pour le codage médico-économique des DPI à des fins statistiques (épidémiologiques) et budgétaires.
La CIM-10, une terminologie de 1re génération (une classification uni-axiale et mono-hiérarchique), est organisée en une hierarchie de 21 chapitres couvrant tout le champ connu à ce jour des maladies, des symptômes, des syndromes, ... Ils sont classés par appareil fonctionnel et associés à une lettre (par exemple : I : "'maladies de l'appareil circulatoire"'). Les chapitres sont toujours au niveau le plus élevé de la hierarchie de cette terminologie, ils sont subdivisés en groupes (ou blocs), eux-mêmes divisés en sous-groupe (sous-blocs), sous-blocs composés (optionnel) de catégories et de sous-catégories englobant le contenu des termes CIM10 (Voir Tableau 2.1) ;

Codes CIM-10	Description	Hiérarchie
A00-B99	Certaines maladies infectieuses et parasitaires	Chapitre
B15-B19	Hépatite virale	Groupe
B15	Hépatite aiguë A	Catégorie
B15.0	Hépatite A avec coma thérapeutique	Sous catégorie

TABLE 2.1 – Exemples de codes CIM-10

– La **classification CCAM** (Classification Commune des Actes Médicaux) [12] : c'est le référentiel destiné à coder les gestes pratiqués par les médecins, gestes techniques dans un premier temps puis, à la suite, les actes intellectuels cliniques. Elle succède, pour les actes techniques, la Nomenclature Générale des Actes Professionnels (NGAP) en secteur libéral et hospitalier, et le Catalogue des Actes Médicaux (CDAM) en milieu hospitalier. Cette classification sert à établir :

+ en médecine libérale et en milieu hospitalier, les honoraires des actes techniques réalisés lors des consultations ;

+ dans les cliniques privées, les honoraires pour les interventions réalisées ;

+ dans les hôpitaux publics et privés, le PMSI et sa tarification des séjours hospitaliers transmis à l'assurance maladie dans le cadre de la T2A.

La CCAM est l'équivalent de Current Procedural Terminology (CPT) aux Etats-Unis. C'est une hiérarchie à héritage simple organisée en 19 chapitres (classement

12. http ://www.ameli.fr/accueil-de-la-ccam/index.php/

par appareil et non par spécialité) scindée en deux parties : la $1^{ère}$ partie pour les actes diagnostiques rangés par technique puis par organe, la $2^{ème}$ pour les actes thérapeutiques classés par organe puis par action. Le Code CCAM correspond à une code à 7 caractères, par exemple "'HH - F - A - 001"' pour coder l'acte thérapeutique : Appendice- Appendicectomie, par abord de la fosse iliaque.

– La **nomenclature LOINC** (Logical Observations Identifiers Names and Codes) [McDonald et al., 2003; Macary, 2007]. La description des analyses biologiques au sein d'un système de gestion des laboratoires (SGL) appelle une nomenclature commune s'appliquant à la demande d'examen, à son traitement et à la restitution du résultat à l'ensemble des professionnels en charge du patient.
Avant 1994, aucun système de codage universel pour précoordonner les noms d'examens biologiques n'existe, bien que des travaux considérables ont été établi au sein des organisations telles que IFCC/IUPAC (International Union of Pure and Applied Chemistry) [13] Committee/Commission on Properties and Units in Clinical Chemistry [Rigg et al., 1995; Olesen, 1996] et Euclides [14]. LOINC a été développé en 1994 par un groupe de chercheurs de l'institut Regenstrief pour répondre à ce besoin [McDonald et al., 2003]. Seule la partie « la classe *laboratoire* »est considérée ici. Selon LOINC, une analyse biologique se décompose en six axes : analyte ou substance mesurée ; propriété mesurée (par exemple : concentration molaire) ; temporalité qui indique si la mesure s'applique à un instant donné ou correspond à une moyenne sur une durée définie ; milieu ou lieu de détermination (sang, urine,...) ; échelle qui indique si la mesure est quantitative, ordinale ou nominale ; technique qui n'est précisée que lorsque des méthodes différentes aboutissent à des résultats différents avec une implication clinique (voir Tableau 2.2).
Le profil IHE « Sharing Laboratory Reports »propose, sans l'imposer, un sous-

Codes LOINC	termes LOINC (subance :propriété :temporalité :milieu :echelle)	Libellé
2951-2	SODIUM : SCNC : PT : SER/PLAS : QN	
2956-1	SODIUM :SRAT : 24H : UR : QN	
1514-9	GLUCOSE 2H POST 100 G GLUCOSE PO : MCNC : PT : SER/PLAS : QN	
14749-6	Glucose : SCNC : PT : SER/PLAS : QN	

TABLE 2.2 – Exemples de codes LOINC

ensemble de la nomenclature internationale LOINC, fruit du travail de sélection

13. http ://www.iupac.org
14. European standard for clinical laboratory data exchange between medical information systems

et de traduction d'un groupe de biologistes français piloté par l'APHP [15] et la SFIL [16] [Macary, 2007].

– La **classification ATC** (Anatomique, Thérapeutique et Chimique) [Skrbo et al., 2004], contrôlée par le "Collaborating Centre for Drug Statistics Methodology" de l'OMS, permet de classer les médicaments en 14 groupes principaux sur la base de l'organe ou du système sur lequel ils agissent ; ensuite ils sont encore répartis sur la base de leurs propriétés chimiques, pharmacologiques et thérapeutiques en quatre niveaux supplémentaires. Un code ATC a la forme générale suivante : **LCCLLCC** (où **L** : lettre ; **C** : chiffre). Dans ce système, les médicaments sont classés en groupes à cinq niveaux différents :
  + 1^{er} niveau : classe anatomique principale (1 caractère alphabétique) ;
  + $2^{ème}$ niveau : sous-classe thérapeutique (2 chiffres) ;
  + $3^{ème}$ niveau : sous-classe pharmacologique (1 caractère alphabétique) ;
  + $4^{ème}$ niveau : sous-classe chimique (1 caractère alphabétique) ;
  + $5^{ème}$ niveau : substance active (2 chiffres).

  A chaque niveau de la classification correspond un code et un libellé ATC. Le libellé du cinquième niveau correspond à la DCI (Dénominations Communes Internationales) [17] de la substance, quand elle existe (voir Tableau 2.3).

Code	Niveau	Libellé	Groupe
A	1	Voies digestives et métabolismes	Groupe anatomique principal
A10	2	Médicaments du diabète	Sous-groupe thérapeutique
A10B	3	Antidiabétiques sauf insulines	Sous-groupe pharmacologique
A10BA	4	Biguanides	Sous-groupe chimique
A10BA02	5	Metformine	Substance chimique

TABLE 2.3 – Exemples de codes ATC de la substance *Metformine*

– **SNOMED CT** (Systematized Nomenclature of Medecine. Clinical Terms) [18] est une terminologie clinique qui fournit un contenu clinique et une expressivité dans

15. http ://www.aphp.fr/

16. Société Française d'Informatique de Laboratoire

17. "Les DCI permettent d'identifier les substances pharmaceutiques ou leurs principes actifs". Directives générales pour la formation de dénominations communes internationales applicables aux substances pharmaceutiques. URL : http ://www.who.int/medicines/services/inn/GeneralprinciplesFr.pdf

18. La SNOMED (nomenclature systématique de médecine humaine et vétérinaire) est à l'origine, une extension à l'ensemble de la médecine du concept développé par le College of American Pathologists [CAP, 2006] avec la SNOP (Nomenclature Systématique d'anatomie Pathologique, 1965). La France a fait le choix de la SNOMED International (SNOMED 3.5). Cette version est incluse dans l'UMLS et plus de 90% des termes de SNOMED International sont présents dans SNOMED CT

le domaine clinique. La manière dont les entités qu'elle représente sont organisées, suggère qu'elle pourrait s'apparenter à ce que l'on désigne sous le terme d'ontologie ou modèle ontologique. Elle est le résultat de la fusion de la SNOMED Reference Terminology (SNOMED RT) et la Clinical Terms version 3 (CT v3) de la "NHS-Service de santé publique du Royaume-Uni". SNOMED est une terminologie multi axiale, organisée en une hiérarchie de 18 catégories de premier niveau : les procédures, les entités observables, les structures du corps humain, les événements, etc. Ces entités, dites majeures, sont regroupées autour d'une racine appelée "Top". SNOMED CT est support à l'indexation, la recherche et le traitement des documents cliniques [Bodenreider et al., 2007].

Au sein de l'équipe CISMeF, nous avons développé plusieurs méthodes conceptuelles, TAL et statistiques pour proposer automatiquement des traductions [Merabti et al., 2009, 2011].

2.3.1.2 Terminologies d'interfaces (TI)

Le codage systématique des données est essentiel pour leur réutilisation, l'analyse et l'interprétation [Cimino, 1998]. Les TR sont très importantes comme support aux applications du SIC [Rector, 1999]. Néanmoins, les professionnels ne sont pas familiers avec les principes de ces TR, très complexes à utiliser et identifiées comme une importante barrière dans l'adoption du DPI [Kamadjeu et al., 2005; Anderson, 2007; Boonstra and Manda, 2010]. C'est pourquoi, il semble plus intéressant de rendre une certaine liberté aux professionnels dans la saisie de données (en les laissant utiliser leur propre vocabuaire) et de prévoir des terminologies adaptées à `ce vocabulaire` dites `TI (ou Terminologie clinique)`. Une revue sur les TI de [Rosenbloom et al., 2006] fournit un aperçu des différentes exigences entre les TI et les TR et la nécessité d'une balance entre la couverture terminologique du domaine et la facilité d'utilisation de cette terminologie par les professionnels.
Les avantages des TI sont divers. Les TI s'adaptent aux besoins informationnels des professionnels en leur fournissant une collection de phrases (termes) se rapprochant du niveau de spécificité connu dans leur pratique de codage [Cimino, 2006; Natarajan et al., 2010]. Les systèmes terminologiques (TR) conçus en Europe et aux Etats-Unis ne sont pas applicables partout et particulièrement en Afrique [Kamadjeu et al., 2005], donc les TI sont une solution pour combler le fossé entre les théories sur les TR de [Cimino, 1998; Rector, 1999] et leur application en Afrique [Kanter et al., 2008]. Par exemple, les TI permettront d'inclure des termes médicaux dans un langage local (voir Tableau 2.4).
Ce point est si important que l'équipe CISMeF a récemment gagné le projet TerSan pour créer des TI en Biologie, en Radiologie, en Soins infirmiers [19] dans le programme TecSan 2011 (ANR). Avant même le début du projet TerSan, une thèse sur ce sujet a débuté en Octobre 2011 avec Nicolas GRIFFON, assistant régional de recherche.

Codes SNOMED CT	Anglais	Fançais	Kinyarwanda	Swahili
271737000	Anemia	Anémie	kubura amaroso	upungufu wa damu
195967001	Asthma	Asthme	Asima	

Table 2.4 – Exemples de termes d'inteface [Kanter et al., 2008]

2.3.1.3 Discussions sur notre positionnement sur les TR/TI

Dans [Dirieh Dibad et al., 2009], nous avons montré que l'objectif de la RI reste le même dans le double contexte (RI orientée "Soins" au sein du DPI et RI orienté "Web"

19. projet en liste d'attente

au sein du portail CISMeF). Nous partageons le même modèle (principe) d'indexation car nous utilisons les mêmes TR pour indexer les données du DPI et les ressources Web dans CISMeF.
Dans ces travaux, [Sakji, 2010] a évalué l'apport de la multi terminologie dans la recherche d'information et a montré une plus-value de 16% en terme de précision. Les travaux de [Merabti et al., 2009] sur les relations intra et inter terminologiques ont permis de mieux répondre à la requête de l'utilisateur et ce via une expansion sémantique ou la reformulation de la requête.
Dans le tableau suivant, nous résumons les contextes d'utilisation de ces systèmes terminologiques pour la recherche d'information, l'exploration du DPI, les revues de la littérature sur ces contextes.

TR/TI	Contexte d'utilisation	Recherche d'information (RI)- Visualisation (V)	Travaux (RI,V)
CIM10	Codage médico-économique et statistique	Non	[Pereira et al., 2006]
SNOMED CT	Codage descriptif	RI, V	[Patel and Cimino, 2007; Pereira et al., 2009]
CCAM	Codage médico-économique	Non	-
LOINC	Codage des analyses biologiques	Non	-
ATC		Non	[Sakji et al., 2009a,b]
TI	Prescription, Codage	RI,V	[Rosenbloom et al., 2006]

TABLE 2.5 – Contexte d'utilisation des terminologies

2.3.2 Normalisation des communications sur les données et entre les systèmes

Le DPI, étant au coeur du système de santé, doit communiquer avec de nombreux systèmes. Les SI médicaux ont été créés sans concertation quant à leur architecture et par conséquent la communication des SI dans un but de suivi médical ou épidémiologique est alors trés difficile. Reconstruire des systèmes tout entier est impossible donc, dans un contexte où les échanges de données sont un véritable enjeu du fait de la complexité des connaissances à échanger et de la variété des acteurs concernés (systèmes, applications, utilisateurs, ...), la question des standards est centrale. Les nouveaux projets comme le Dossier Médical Personnel (DMP) et la mise en place des dossiers de spécialités partagés comme le Dossier Communicant en Cancérologie [20](DCC) amplifient le problème et la nécessité de l'intégration des systèmes hétérogènes car ils impliquent le développement d'échanges de connaissances. Il est donc impératif que chaque projet ne génère pas ses propres normes et formats mais que les systèmes tendent vers des composants communs. Les autorités médicales préconisent donc désormais le passage d'une logique d'interface point à point, à une logique de partage d'informations entre les systèmes, en vue de l'accomplissement du processus de soins. Cette échange d'information n'est possible que si une normalisation est retenue par l'ensemble des acteurs du système de santé, c'est ce qu'on appelle l'interopérabilité [Eichelberg et al., 2005]. C'est une solution parmi le panorama d'approches existantes (approche ontologique [Diallo, 2006], approche par médiateur Beneventano et al. [2000], approche par les web services Hansen et al. [2003]; Murphy et al. [2006], ...) pour répondre aux problèmes de cloisonnement des SIS.
Cette approche se tourne vers la construction d'architectures standards destinées à faire coopérer des systèmes hétérogènes. Des travaux existent dans ce domaine mais nous nous limitons `aux normes standards`. Des travaux sont en cours pour construire et pour diffuser ces normes.
En santé, les principaux organismes influents sont :

- au niveau national : EDISanté [21], AFNOR [22], HPRIM (HL7 français) [23] ;
- au niveau européen : CEN/TC251 [24] ;

20. http ://www.e-cancer.fr/soins/parcours-de-soins/dossier-communicant-de-cancerologie
21. http ://www.edisante.org/
22. http ://www.afnor.fr
23. http ://www.hprim.org
24. http ://www.centc251.org

– au niveau international : ISO/TC 215 [25], DICOM [26], HL7 [27], et IHE [28].

Ce tableau 2.6 représente une liste non exhaustive des standards de communication existants en les classant en 2 catégories : ceux spécifiques à la santé et les autres génériques. Nous présentons dans les paragraphes suivants deux standards utilisés dans la plupart

Standards spécifiques	Standards génériques
Health Level Seven (HL7)	eXtensible markup Langage (XML) [29]
HL7 Clinical Document Architecture (HL7 CDA)	Resource Description Framework (RDF) [30]
Cross-Enterprise Document Sharing (XDS)	Simple Object Access Protocol (SOAP) [31]
Digital Image COmmunication in Medecine (DICOM)	...
Medical Markup Language (MML) [32]	

TABLE 2.6 – Liste non exhaustive des standards

des cas pour l'échange et la partage des données du DPI.

2.3.2.1 HL7/CDA

L'objectif de l'organisation HL7 est de créer des standards flexibles et peu coûteux, des méthodologies permettant l'interopérabilité des SIS et le partage des DPI. Pour cela, le protocole HL7 fournit un vocabulaire et une grammaire dédiés au domaine de la santé, permettant d'exprimer des données médicales sous la forme de messages « ayant du sens » pour les applications. La version 3 d'HL7 est un standard de messagerie basé sur XML et structuré à partir d'un modèle objet, le RIM, permettant ainsi l'application de spécifications comme le CDA.
CDA [Dolin et al., 2001, 2006] est un standard ANSI depuis Mai 2005. Il s'agit d'un marquage des documents médicaux qui spécifie, dans une perspective d'échange, la structure et la sémantique des documents cliniques. Ce standard est compatible avec XML et le RIM. Un document CDA est un assemblage persistant d'informations, légal, authentifiable et appréhendable uniquement dans son ensemble et dans son contexte d'utilisation par des humains. Un tel document peut inclure des textes, des images, des sons ou tout autre contenu multimédia [Dolin et al., 2006]. La revue d' [Eichelberg et al., 2005] décrit

25. http ://www.iso.org/iso/en/stdsdevelopment/tc/tclist/TechnicalCommitteeDetailPage.TechnicalCommitteeI
=4720
26. http ://medical.nema.org/
27. http ://www.hl7.org
28. http ://www.ihe.net/

de manière détaillée ce standard. En somme, un document CDA comporte un en-tête et un corps :

- l'en-tête donne le contexte du document (identification du document, Signataire, Destinataire(s), Auteur(s), ...) pour faciliter les échanges et la gestion des documents ;
- le corps comporte des informations cliniques selon les 3 niveaux de description définis par HL7 :
  + 1^{er} niveau : un bloc de texte codé en base64 ;
  + $2^{ème}$ niveau : un bloc de texte narratif *(Section-Level)* ;
  + $3^{ème}$ niveau : un bloc de texte narratif et des blocs de données codées et structurées *(Entry-Level)* pour faciliter le traitement informatique des blocs de texte.

Dans la figure 2.3, HL7/CDA fait partie du cadre ontologique puisqu'il fournit une sémantique pour donner du sens aux données.

2.3.2.2 IHE/XDS

L'Integrating Healthcare Enterprises (IHE) [33] est une approche innovante fondée sur une coopération étroite entre les clients et les industriels. L'IHE est lancée en 1998 aux US sous l'impulsion de RNSA (Radiological Society of North America) et d'HIMSS (Healthcare Information and Management Systems Society). L'objectif principal d'IHE est de développer les échanges et l'intégration entre logiciels de différents fournisseurs en s'appuyant sur des normes et des standards (par exemple HL7) reconnus et opérationnels afin de résoudre des problèmes d'interopérabilité et d'intégration non encore résolus en l'état actuel. En 2000, la France prend le leadership européen sur la démarche IHE avec la collaboration entre la Société Française de Radiologie (SFR) et le Groupement pour la Modernisation des Systèmes d'Information Hospitalier (GMSIH).

L'approche IHE, pour une meilleure interopérabilité, est de développer des interfaces point à point permettant à une application émettrice de transmettre à une application destinatrice un message qui soit lisible *(format du message)*, compréhensible *(sémantique du message)* selon un protocole de communication commun *(transport du message)*. C'est une approche pour donner "un sens" aux fonctionnalités des applications. Pour cela, des profils d'intégration ont été créés (21 en 2004, 72 en 2007) pour spécifier les différentes **transactions IHE** (échange d'information) possibles entre les **acteurs IHE** (SI, applications,...) pour réaliser une tâche, c'est-à-dire les éléments dont on a besoin pour intégrer un composant logiciel (trigger d'événements, messages, données).

33. Que l'on peut traduire par "Résoudre les problèmes d'intégration en s'appuyant sur les standards"

L'un de ces profils d'intégration se nomme `XDS`. Ce dernier permet de gérer le partage de documents cliniques entre les acteurs de santé. XDS propose une approche pragmatique intégrant à la fois la perspective CDA sur lequel il base la définition de ses méta-données et ebXML en terme de communication et d'architecture de services. Ce standard permet à différents DPI (hôpitaux, réseaux de soins, ...) de partager des documents médicaux. Il sert de base à de nombreux projets, dont le DMP.
Dans la figure 2.3, XDS fait partie du cadre d'application puisqu'il fournit une sémantique pour donner du sens aux fonctionnalités des applications.

2.3.2.3 Entre Intégration et Interopérabilité

Le standard ISO 14258 ISO [1998] a clarifié la différence entre l'intégration et l'interopérabilité. ISO 14258 considère que des modèles peuvent faciliter :

1. l'intégration quand il existe un standard qui les formalise HL7 [2005] ;
2. l'unification quand il existe un méta-modèle pour assurer une équivalence sémantique ;
3. la fédération lorsque les modèles existent par eux-mêmes mais que les correspondances entre les concepts qu'ils modélisent peuvent être définies à un niveau ontologique pour formaliser la sémantique de leur interopéarabilité [Diallo, 2006].

L'interopérabilité est un moyen d'assurer l'intégration des systèmes selon quatre niveaux (voir figure 2.3). Nous ne détaillons pas les problématiques sur l'intégration et/ou l'interopérabilité des systèmes. Nous indiquons qu'un bon modèle du DPI doit prendre en compte certains niveaux d'intégration pour englober les différents types de données du DPI et ainsi de le rendre flexible et réutilisable dans un autre système (cf. paragraphe 4.3.4.2 du chapitre 4).
Pour résumer, [Lenz et al., 2007] ont présenté une matrice pour classifier ces différents standards selon deux perspectives selon : **l'objet de l'intégration** en distinguant deux niveaux d'intégration (*"Intégration des données"* et *"Intégration fonctionnelle"* et **la portée des standards** en distinguant deux autres niveaux (*"Intégration technique"* et *"Intégration sémantique.*
Dans la figure 2.3, nous avons décrit ces 4 niveaux et nous classifions ces standards en distinguant les standards existants de ceux utilisés dans notre travail.

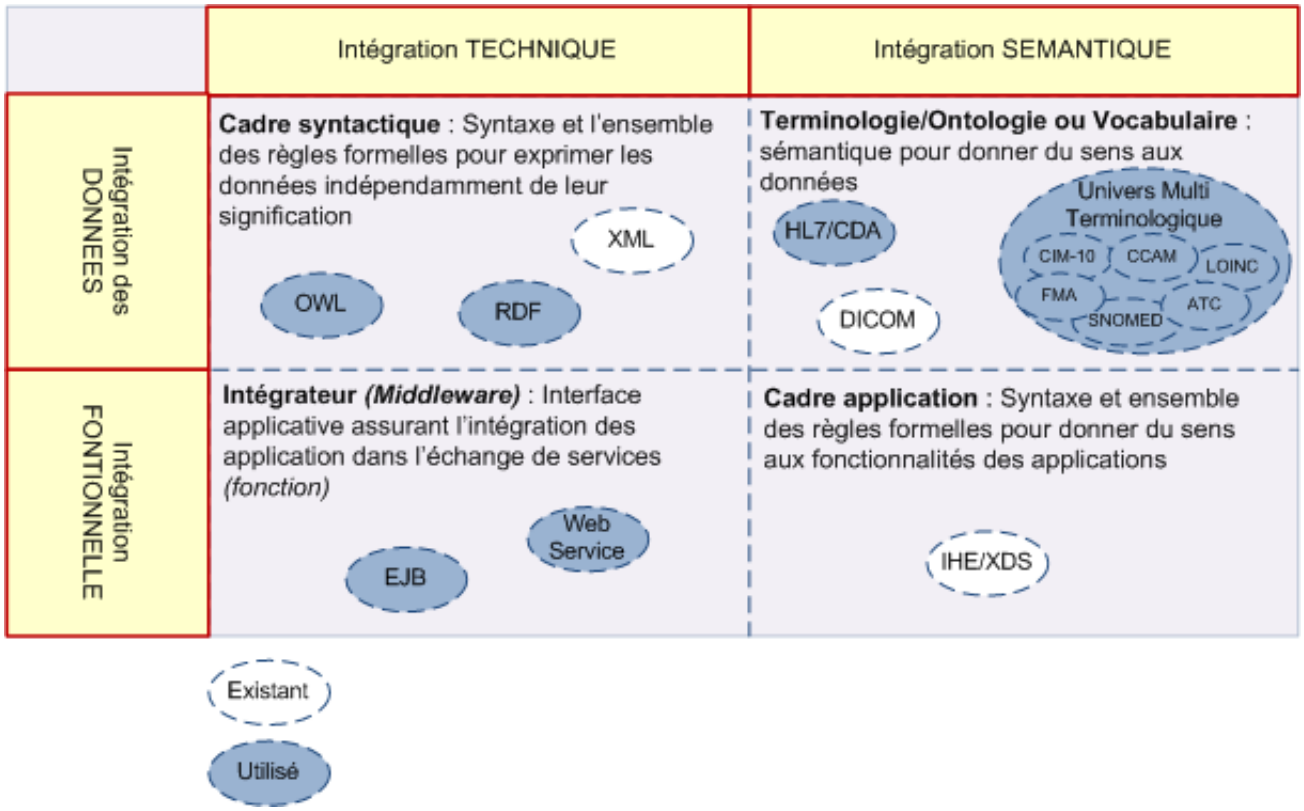

FIGURE 2.3 – Classification des différents standards adaptée de Lenz et *et al.*,(2005)

Synthèse

Dès le début du développement de l'informatique médicale, l'un des facteurs les plus critiques est la qualité des données médicales. Plusieurs tentatives ont été entreprises pour encourager les professionnels à utiliser les systèmes informatiques.
Ce chapitre nous a permis d'aborder les problématiques de modélisation de ces données sous une forme à la fois *structurée* selon des modèles d'information [de données] et *standardisée* à l'aide de vocabulaires standards, des moyens dont la seule fin est l'amélioration de la qualité des données et de son réutilisation.
De nombreux travaux ont été consacrés à cette modélisation à différentes fins (RI clinique/épidémiologique, accès à l'information, codage, échange et partage, ...). Les travaux sur la modélisation du DPI ont permis d'aboutir à différents modèles standards [et non standards]. Ces modèles permettent d'utiliser des outils standards (HL7/CDA, IHE/XDS, Web Sémantique, Archétypes) pour raisonner sur les données pour la plupart et pour partager et communiquer pour d'autres comme les modèles RIM HL7 et HISA. Ces derniers sur lesquels sont basées au moins pour partie, la plupart des applications récentes de gestion du DPI, ne sont pas adaptés à la RI du fait des nombreuses tables contenant des données nécessitant des requêtes complexes comme nous avons pu l'observer sur un modèle pourtant simplifié de l'application de gestion des dossiers patients du CHU de Rouen. Nous évaluerons certains modèles en détaillant leurs caractéristiques conceptuelles et techniques et nous comparerons ces modèles avec le modèle proposé respectivement dans les sections 4.2.2 et 4.4 du chapitre 4.
Nous allons voir dans le chapitre suivant, les intérêts et les problématiques de la RI qui ont amené à concevoir des modèles d'information [de données] dans un contexte de soins (le DPI).

CHAPITRE 3

La Recherche d'Information

Introduction

Actuellement, les applications informatiques de gestion des dossiers patients, privilégient les fonctions d'enregistrement et de communication. Elles ne disposent pas de fonctions avancées de recherche et d'extraction d'information, ce qui, paradoxalement, complique parfois la tâche par rapport au dossier papier [Thomas et al., 2006]. On est loin de l'objectif théorique du DPI : mettre les informations concernant un patient à disposition des professionnels de santé où et quand ils en ont besoin [Ondo et al., 2002]. Dans la pratique quotidienne du médecin, ces fonctions sont source de gain de temps lors de la prise en charge, en fournissant les informations en temps voulu. Par ailleurs, ces fonctions peuvent être le support d'outils permettant de réduire les erreurs médicales [Ammenwerth et al., 2011], qui peuvent être utilisées pour la prévention et la prédiction de maladies en particulier dans le cadre de la génomique [Takai-Igarashi et al., 2011], faciliter les études cliniques, épidémiologiques et biomédicales [Daniel et al., 2009]. Elles peuvent aussi être partie intégrante d'outils d'aide à la décision et d'évaluation des données du DPI [Renaud-Salis et al., 2010].
Disposer d'outils de RI est un besoin, afin de combler le fossé ''*sémantique*'' entre les besoins informationnels de l'utilisateur et les éléments d'information du DPI. Ce chapitre va faire un état de l'art sur les approches des systèmes de recherche d'information dans le domaine médical (voir Figure 3.1).

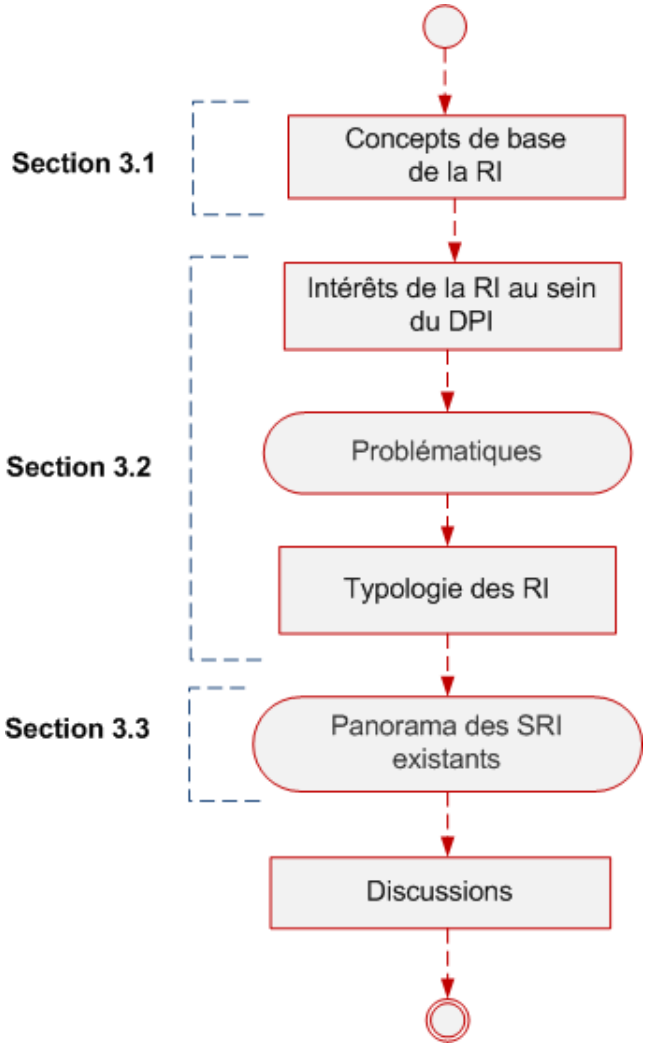

FIGURE 3.1 – Schéma synoptique de notre recherche d'information

3.1 Concepts de bases de la RI

3.1.1 Donnée versus Information

Les termes *"donnée"* et *"information"* sont fréquemment utilisés dans le domaine de recherche et de traitement. Il y a un manque de différenciation entre ces deux termes. Dans cette section, nous soulignerons cette différence en comparant la différence entre *la recherche d'information (RI)* et *la recherche de données (RD)*.

Pour [Baeza-Yates et al., 1999], la RD (*data retrieval*) consiste à fournir un ensemble de données [1] bien déterminé à travers une requête bien définie et complète. Le but d'un système orienté *RD* est de satisfaire exactement la requête. A l'inverse, la RI (*information retrieval*) utilise le traitement de la langue pour répondre au besoin informationnel de l'utilisateur pour retrouver des éléments d'informations. Un système orienté *RI* tente d'interpréter la sémantique de l'information en la classant selon la pertinence de la demande d'information.
Le tableau 3.1 représente la différence entre ces deux types de recherche.

	Recherche de données (RD	**Recherche d'information (RI)**
Alignement	Alignement exact	Alignement partiel ou la plus proche
Raisonnement	Déductif	Inductif
Modèle	Déterministe	Probabiliste
Langage de requêtes	Simple	Traitement de langue
Spécification de requêtes	Complète	Incomplète
Recherche	Correspondance	Pertinence
Erreurs	Avec conséquence	Sans conséquence

TABLE 3.1 – Différences entre une RD et une RI

3.1.2 Une brève description des systèmes de recherche d'information (SRI)

Par définition, un SRI a pour fonction de sélectionner dans une collection de documents ceux qui sont susceptibles de contenir les informations dont a besoin l'utilisateur.

1. *"une donnée"* est un élément atomique avec une interprétation bien définie

Son but est de retourner à ce dernier le maximum de documents pertinents répondant à son besoin et le minimum de documents non pertinents [Sakji, 2010].
Selon [Koopman et al., 2011], les SRI permettent de combler le fossé sémantique entre le besoin informationnel de l'utilisateur et le contenu des documents (les dossiers médicaux).
Nous ne détaillons pas les SRI dans cette thèse car ils ont déjà été abordés plus en détail [Sakji, 2010].
Nous faisons une synthèse des différents processus de la RI selon [Baeza-Yates et al., 1999] en 4 modules (Figure 3.2) :

- un module **D** qui comprend la base de données stockant le corpus de documents indexés. Un document consiste en un ensemble d'éléments d'information décrits par des métadonnées [Hersh, 2008] (cf . sous section 3.1.2.1) ;
- un module **O** qui comprend les différentes opérations de traitement sur les documents et les requêtes de l'utilisateur (cf . sous section 3.1.2.2) ;
- un module **R** qui comprend les différents processus de la RI pour rechercher et retourner à l'utilisateur un ensemble de documents correspondants au mieux à la pertinence-utilisateur et la pertinence-système (cf . sous section 3.1.2.3) ;
- un module **U** qui comprend les interfaces de RI par lesquelles l'utilisateur [re]formule sa requête et visualise les résultats de sa recherche.

3.1.2.1 Les métadonnées

Par définition, une métadonnée est *"une information structurée qui décrit, explique ou qui rend plus facile la recherche, l'utilisation et la gestion d'éléments d'information"* [Niso, 2004].
Les métadonnées sont créées pour répondre à différentes fonctionnalités parmi lesquelles l'amélioration de la RI [Niso, 2004]. Ces métadonnées ont pour objectif de caractériser le document et couvrir au mieux le contenu sémantique du document.
Les termes d'indexation représentent la sémantique du document (son contenu). Ils varient de mots, qui apparaissent dans le document à des termes spécifiques assignés par des indexeurs professionnels. Ces métadonnées sont le résultat de l'indexation de documents en utilisant les terminologies de références. Pour le catalogue CISMeF, les métadonnées, pour décrire les caractéristiques d'un document Web [Thirion et al., 2004], sont essentiellement issues du Dublin Core [Darmoni et al., 2001a], et les termes d'indexation sont pour la plupart issus du thésaurus MeSH.

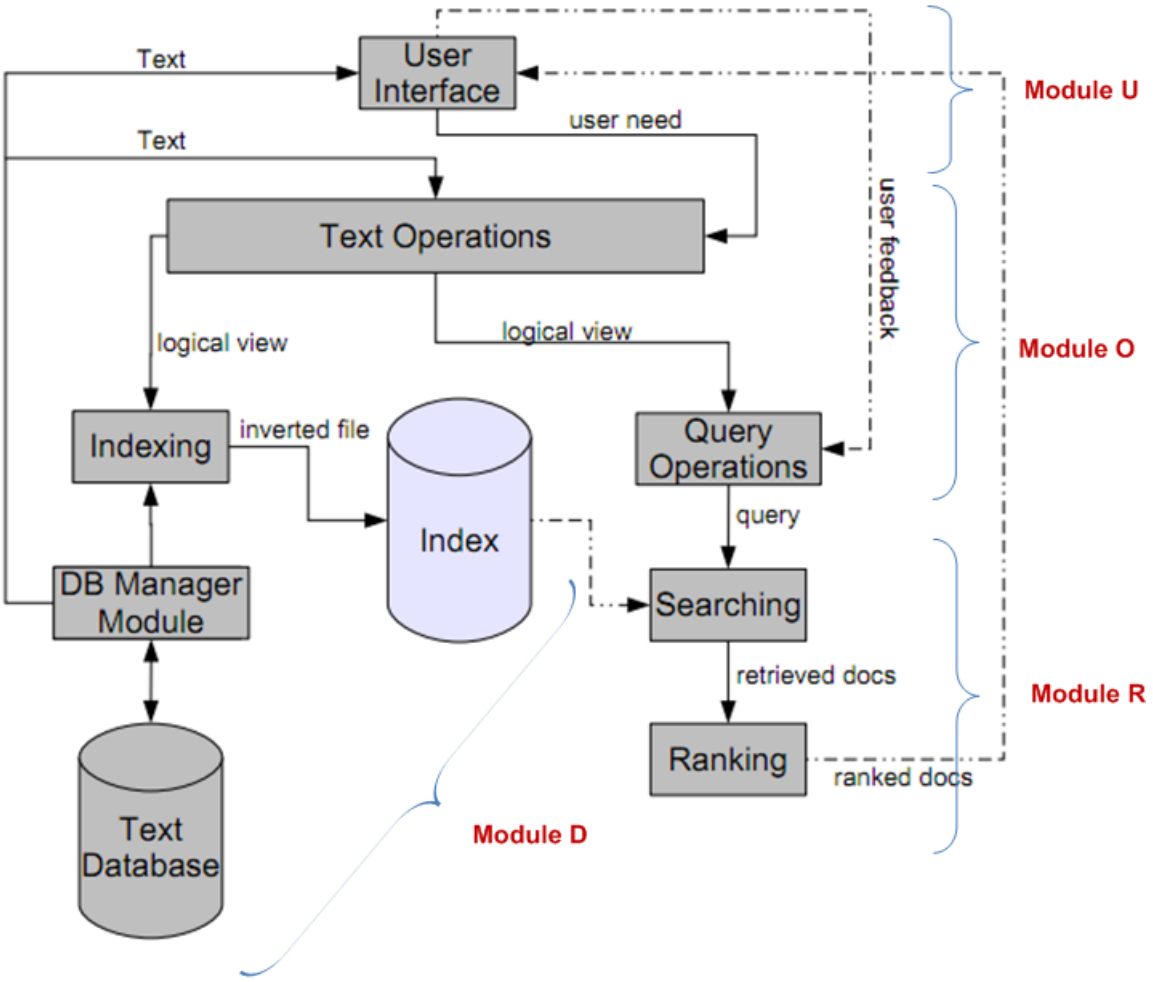

FIGURE 3.2 – Schéma synoptique de la RI [Baeza-Yates et al., 1999]

3.1.2.2 Les traitements des documents et des requêtes

En ce qui concerne les documents, l'indexation automatique [2] repose sur des algorithmes associant automatiquement des métadonnées à des parties de documents. Plusieurs techniques (*désuffixation, lemmatisation, ...*) sont utilisées et dépendent du SRI. La figure 3.3 résume les différentes opérations qui peuvent être effectuées sur un document pour obtenir l'ensemble des métadonnées.

Les principales limites de l'indexation automatique sont les algorithmes qui exploitent

2. Ils existent d'autres types d'indexation [Sakji, 2010] : l'indexation manuelle effectuée par les indexeurs professionnels, l'indexation supervisée qui est une validation de l'indexation automatique par un indexeur professionnel, ...

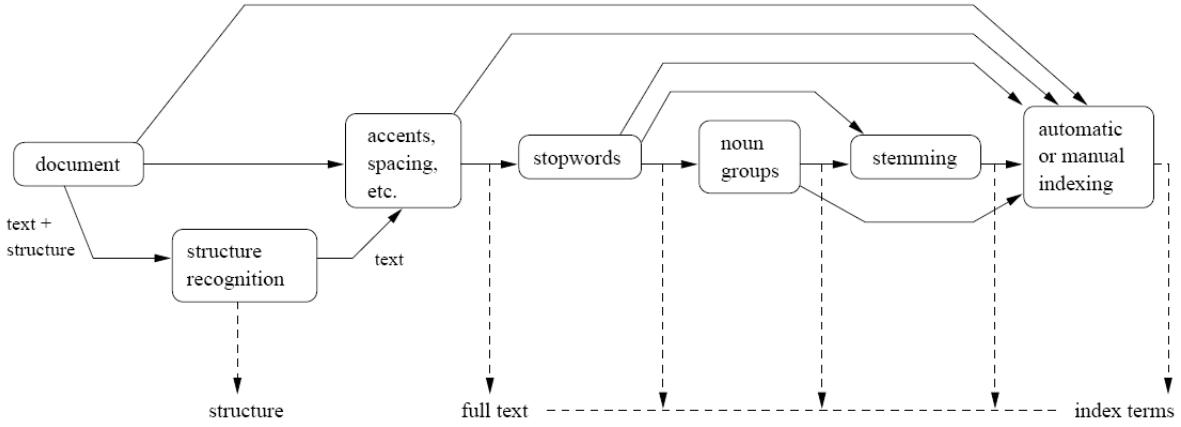

FIGURE 3.3 – Schéma synoptique des différentes phases de [pré]traitement d'un document [Baeza-Yates et al., 1999]

l'information contenue, dans les documents alors que leur interprétation est dépendante de l'information contextuelle accessible hors des documents [Sakji, 2010].
Une fois que les métadonnées sont créées pour un document et stockées dans le corpus d'index, des traitements sont exécutés pour reformuler la requête de l'utilisateur en une requête machine afin que le SRI retourne un ensemble de documents pertinents. Une des problématiques sous jacente est comment améliorer cette pertinence.

3.1.2.3 Amélioration de la RI

Plusieurs approches existent pour améliorer la RI et donc, améliorer la pertinence des documents retrouvés :

– une approche *orientée système* dont l'objectif est de mieux comprendre la sémantique d'un document afin d'améliorer le processus d'indexation et de RI.

  Plusieurs techniques existent, parmi lesquelles : (1) les méthodes de pondération des termes d'indexation et/ou de la requête, et (2) les méthodes d'expansion sémantique des requêtes en exploitant les relations sémantiques ou le feedback de l'utilisateur ;

  (1) cette méthode associe un poids à chaque terme indexé et permet de contribuer à la différentiation informationnelle des documents [Sakji, 2010]. Les formules de pondération les plus connues sont **TF*IDF** et sa variante **okapi BM25** [Hersh, 2008] ;

(2) cette méthode consiste à étendre sémantiquement la requête utilisateur en exploitant les relations hiérarchiques et sémantiques des termes d'indexation.

– une approche *orientée utilisateur* dont l'objectif est d'aider l'utilisateur dans sa recherche.

L'utilisateur examine la liste des documents retrouvés et identifie ceux qui sont pertinents. Le feedback de l'utilisateur permet la reformulation de la requête avec de nouvelles informations. [Tamine et al., 2007] mettent en évidence l'adaptation du cycle de vie d'un processus d'accès à l'information, à un utilisateur spécifique, en vue de lui délivrer une information pertinente par rapport à ses besoins, son contexte et ses préférences.

3.1.3 Évaluation des SRI

L'évaluation d'un SRI consiste à mesurer la différence entre un résultat obtenu et un résultat attendu. Au dèlà du temps de réponse et de l'espace mémoire consommé pour stocker un corpus de documents, la pertinence des résultats obtenus est une mesure pour évaluer la performance d'un SRI.
Cette évaluation a un double but : (1) comparer les performances des différents SRI, (2) analyser l'évolution de cette performance en fonction de l'évolution du SRI (nouvelle méthode ou algorithme de RI).

3.1.3.1 Les mesures d'évaluation

Pour analyser l'évolution de la performance d'un SRI, deux types d'évaluation existent [Hersh, 2008] :

– une évaluation *orientée sur le système* :

Plusieurs approches existent [Sakji, 2010], la plus usuelle est l'utilisation du *rappel* **(R)** et de la *précision* **(P)**.
Nous supposons, pour la requête d'un utilisateur, **(D)** le nombre de documents pertinents, **(D')** le nombre de documents trouvés et **(D")** le nombre de documents retrouvés pertinents.
Le *rappel* (**R=D"/D**) est la capacité d'un SRI à ramener tous les documents pertinents (`completeness`) et la *précision* (**P=D"/D'**) est la capacité d'un système à ne sélectionner que des documents pertinents (`relevance`) [Sakji, 2010].
L'augmentation de l'un induit le plus souvent une diminution de l'autre. Il est néces-

saire de faire un compromis entre ces deux valeurs. Un SRI est jugé performant s'il réussit à trouver l'équilibre, le juste milieu, entre ces deux critères de mesure (**P et R**). Ce compromis semble dépendant du cas d'utilisation et implique la combinaison. D'où la *F-mesure* **(F)** [3] qui est une moyenne pondérée de la précision et du rappel (**F** = 2***P*R**/**P**+**R**).

– un évaluation *orientée sur l'utilisateur* :

L'approche usuelle est l'utilisation de la *couverture* et la *nouveauté*.
Nous supposons, pour la requête d'un utilisateur, **(U)** le nombre de documents pertinents, **(U')** le nombre de documents trouvés, **(A)** le nombre de documents connus par l'utilisateur, et **(A')** le nombre de documents pertinents non connus par l'utilisateur.
La *couverture* (**C**= **A**∩**U**/**A**) est la capacité d'un SRI à ramener tous les documents pertinents et connus par l'utilisateur et la *nouveauté* (**N**=**A'**/**A'**+**A**) est la capacité d'un SRI à ramener des documents pertinents et non connus par l'utilisateur.

3.1.3.2 Les campagnes d'évaluation

Pour comparer des SRI par leur performance, des campagnes d'évaluation ont vu le jour. Un des objectifs des campagnes d'évaluation est d'évaluer et de mesurer l'efficacité des systèmes de recherche d'information, développer la communication entre l'industrie, l'université et l'état, en mettant en place un forum ouvert pour faciliter les échanges d'idées sur la recherche [Sakji, 2010].
Parmi les projets les plus ambitieux dans ce cadre, nous pouvons citer les campagnes d'évaluation de CLEF (Cross Language Evaluation Forum) [4] qui ont pour objectif de promouvoir la recherche et le développement dans le domaine de la recherche d'information multilingue, d'une part en offrant une infrastructure pour tester et évaluer les SRI sur des supports écrits dans les différentes langues européennes, en mode monolingue, multilingue ou interlangue, et d'autre part en mettant au point des séries de tests composés de données qui peuvent être réutilisées par les développeurs de systèmes, pour l'évaluation.
A notre connaissance, il n'y a pas un framework standard d'évaluation d'outils de RI dédié aux dossiers médicaux. Dans la campagne de CLEF, nous avons **ImageCLEFmed** qui permet l'évaluation de la performance des SRI médicaux, fondée sur des collections d'images décrites en mono ou multilingues [Müller et al., 2008].

3. Il s'agit d'un cas particulier de la mesure générale F_β de []
4. URL : http ://www.clef-campaign.org/

[Koopman et al., 2011] proposent un corpus test avec des requêtes types et avec leur *"relevance judgement"* spécifique au dossier médical [5].

5. Medical Information Retrieval Test Collection URL : http ://aehrc.com/med_eval/

3.2 Recherche d'information au sein d'un DPI : Intérêts, Problématiques et Typologies

3.2.1 Intérêts de la RI pour le DPI

L'enjeu de la RI pour le DPI est l'amélioration de la qualité des soins [Hersh, 2008]. Améliorer la qualité de soins implique intrinsèquement d'améliorer les outils de travail du soignant.

3.2.1.1 Du processus de soin ...

Dans un contexte de soin, l'objectif de la RI consiste à procurer aux médecins l'information requise au bon endroit, au bon moment et à la bonne personne [Ondo et al., 2002]. Des études ont montré, que dans les unités de soins ambulatoires, plus de 80% des décisions cliniques étaient retardées ou fondées sur une information incomplète car les informations cliniques importantes ne sont pas à la portée des cliniciens au moment de la consultation. D'où l'intérêt de créer un résumé de l'histoire du patient sous forme de vue adaptée en fonction du profil de l'utilisateur et du contexte.
Les médecins ne font que rarement usage du DPI pour une tâche autre que la consultation d'information [Christensen and Grimsmo, 2008]. L'utilisation de ce type d'outil de RI dans la pratique allège la lourdeur des routines des tâches médicales et offre plus de temps aux médecins pour les soins du patient et donc contribue à améliorer la qualité des soins [Barnett, 2006]. Ces outils peuvent être intégrés dans les SIH afin de rechercher dans des DPI, ils sont différents d'un point de vue typologie et organisation des données : des documents textes pour **MIRS** [Spat, 2007] et **R-oogle** [Cuggia et al., 2010, 2011] ou des documents CDA HL7 pour **XOntoRannk** [Farfan et al., 2009].
La RI au sein d'un DPI est indissociable de la RI orienté-Web. Cette dernière permet d'améliorer les connaissances des soignants. Les travaux de [Cao et al., 2010] ont montré que la recherche pertinente de documents de la littérature médicale spécifique aux patients améliore la qualité des soins.
Le DPI représente un potentiel bénéfice car il constitue une nouvelle interface entre le domaine du soin et de la recherche, pouvant améliorer la portée et l'efficacité de la recherche [Powell and Buchan, 2005] avec des données précises et de qualité [Sanderson et al., 2004].

3.2.1.2 .... au processus de recherche

Parallèlement à l'utilisation accrue du DPI pour les soins cliniques, on a assisté au développement d'outils dédiés pour le recueil des données médicales pour la recherche et les essais cliniques [Collen, 1990]. L'Europe a investi 16 millions d'euros sur 4 ans pour le projet EHR4HC [6], projet qui vise à faciliter l'utilisation des données des dossiers médicaux pour la recherche clinique. Un pan de ce projet consiste à permettre de rechercher les patients qui répondent aux critères d'inclusion et d'exclusion des essais cliniques au sein de plusieurs systèmes d'information médicaux. En effet, l'une des difficultés de la recherche clinique est l'inclusion d'un nombre suffisant de participants dans un temps limité [Schulz et al., 2008].
La constitution de cohortes peut être considérée comme une application particulière de la RI. Les outils de RI multi-patients, permettent de sélectionner des patients répondant aux critères d'inclusion et d'exclusion d'essais cliniques. Souvent les outils existants sont réservés aux spécialistes du domaine [Rokach et al., 2004; Plaza and Díaz, 2010]. Les médecins sont demandeurs de tels outils, l'enquête de [Natarajan et al., 2010] a révélé que certains médecins essayaient d'identifier les patients respectant les critères d'inclusion de leur projet de recherche, en répétant plusieurs fois les mêmes requêtes sur plusieurs dossiers patients.
Dans une étude sur les avis des médecins sur l'utilisation de systèmes de recrutement de patients éligibles pour des essais cliniques fondés sur le DPI [7], la majorité d'entre eux ont trouvé très bénéfique ces types de systèmes, les limites de leur utilisation sont : un manque de temps, un manque d'information sur l'inéligibilité d'un patient ou tout simplement un manque d'information sur ces systèmes [Embi et al., 2008].
Les outils existants sont nombreux comme **I2B2** ou **R-oogle**, **STRIDE** un outil ayant pour vocation de faciliter la recherche translationnelle [Lowe et al., 2009], palliant difficilement l'absence d'outil de recherche multi-patients.
Ces outils de RI multi-patients peuvent s'avérer utiles dans la RI épidémiologique et à la mesure de la qualité.

6. EHR4HC : Electronic Health Records for Clinical Research, URL : http ://www.ehr4cr.eu/
7. EHR-based Clinical Trial Alert system

3.2.2 Problématiques

Rechercher une information médicale pertinente dans un DPI est une tâche plus difficile que de rechercher une ressource web, car plusieurs types de données sont liés aux problèmes de santé d'un patient. Tous ces éléments doivent être reconnus "*pertinents*" même si le besoin informationnel du professionnel n'est spécifique à aucune de ces données. A partir d'éléments tels que : le traitement du patient pour son problème de coeur d'il y a 5 ans, sa tension artérielle, son comportement actuel de fumeur, le médecin définira le bon traitement pour le prévenir d'une future attaque cardiaque. Toutefois, la RI dans le domaine du soin présente plus de similitudes que de différences avec la RI documentaire [Dirieh Dibad et al., 2009]. Nous détaillerons 4 problématiques à prendre en compte dans le développement de SRI orientés *soin*

3.2.2.1 Besoins informationnels *(questions cliniques)*

Pour [González et al., 2007], un professionnel a besoin d'informations de différentes natures pour finir une tâche (une consultation) :

- des besoins informationnels sur le patient à travers son DPI ("*Les informations ophtalmologiques dans le dossier d'un patient suivi depuis longtemps pour un diabète*") ;
- des besoins informationnels sur les connaissances du domaine ("*Quelles sont les recommandations pour la prise en charge d'une adénopathie cervicale ?*").

Dans l'étude d' [Hersh and Hickam, 1998], 22% des questions cliniques (concernant généralement les expériences et les observations du domaine) ont permis de reconnaître ou de diagnostiquer un problème médical. [González et al., 2007] confirment cette étude et ajoutent que 10% des RI sont exécutées durant la consultation. [Natarajan et al., 2010] ont analysé les fichiers logs de leur moteur de recherche **CISearch** pour déterminer les types de RI sur les DPI : sur les 85% des requêtes spécifiques au dossier patient, 29% des requêtes portent sur des données biologiques et 22% sur des maladies et/ou des symptômes.
En ce qui concerne les besoins informationnels inhérents aux connaissances du domaine, plusieurs approches et outils existent : des moteurs de recherche ont été développés comme Doc'CISMeF [Darmoni et al., 2001b], certains proposent un ensemble de questions cliniques génériques générées à travers des interviews [Braun, 2008] ou à travers les données du DPI [Braun et al., 2007], d'autres proposent des systèmes Questions-Réponses(QR) comme [Cao et al., 2010] ou des infos-bulles intégrés aux DPI [Cimino et al., 1992]. A cet effet, une thèse a débuté (Zied Moalla) sur l'application de l'approche QR au moteur de recherche Doc'CISMeF.

Cependant, la solution consistant à définir un ensemble de besoins informationnels génériques pour le cas du DPI est impossible. Car, dans un contexte de soins, le besoin informationnel d'un soignant est partiel [8] et incertain [9] [Gardner, 1997].
Par ailleurs, il n'est pas évident de reproduire les pratiques de RI effectuées sur le dossier patient : première lecture, recherche de faits, résolution de problèmes, etc.
En ce qui concerne les besoins informationnels spécifiques aux données médicales, la solution est le développement d'outils de RI qui permettraient de faciliter grandement deux des trois cas d'utilisation du dossier médical identifiés par [Nygren and Henriksson, 1992] : *"rechercher des détails précis, explorer des hypothèses* [10] *"*.
Deux approches existent pour ces outils de RI :

- une approche orientée *système* qui se concentre principalement sur les données et documents : amélioration de l'indexation des données et/ou de la RI ;
- une approche orientée *utilisateur* qui se focalise sur l'utilisateur pour la (re)formulation de sa requête.

3.2.2.2 Indexation des données & documents

Les outils de RI orientés *système* implémentent différents types de méthodes lexicales, TAL ou sémantiques [Merabti, 2010] pour identifier un ensemble de termes (ou des concepts) *représentatifs* du contenu textuel et pondérés pour différencier leur importance informationnelle [Sakji, 2010].
Dans le domaine du DPI, les travaux sur l'exploitation des données textuelles ont porté sur la reconnaissance de catégories sémantiques et/ou sur la détection de relations sémantiques entre concepts et dernièrement, sur l'indexation et l'extraction des données dans les documents médicaux [Hersh, 2008].
Cependant, les méthodes TAL et d'indexation sémantique des documents médicaux non structurés doivent prendre en compte plusieurs dimensions [Currie et al., 2001] parmi lesquelles, les plus importantes sont : *le contexte, la négation linguistique, l'extraction de valeurs cliniques, ...* Sans prendre en compte le contexte dont laquelle une observation clinique a été produite, ni pouvoir extraire les prescriptions médicamenteuses ou la valeur d'un score d'APGAR, ni différencier des affirmations négatives des assertions positives, il est difficile d'avoir une compréhension pragmatique du texte médical et, du coup, de répondre intelligement aux requêtes utilisateurs.
Ces dimensions sont toujours d'actualité et c'est la raison d'être, des compétitions d'éva-

8. Ce besoin informationnel est incomplet du fait de sa spécifité dans le contexte du patient
9. Ce besoin informationnel est imprécis car le professionnel n'a pas accès à toutes les informations
10. "to search for specific details, and to prompt or explore hypotheses"

luation [11] : CLEF [12], I2B2 [13], TREC [14].

3.2.2.2.1 La temporalité de l'information médicale

De nombreuses approches ont été développées pour exploiter cette dimension. [Zhou and Hripcsak, 2007] les classent en trois catégories :

1. les approches pour répondre aux besoins des SI cliniques. La temporalité est exploitée à travers des vues chronologiques des éléments du DPI [Rogers et al., 2006] ;
2. les approches pour résoudre les problèmes d'incertitudes et de granularité ;
3. les approches pour modéliser le raisonnement sur les données temporelles [Hripcsak et al., 2005].

La temporalité est très importante dans le DPI [Hripcsak et al., 2005]. La dimension temporelle concerne les différents types de données du DPI, autant les données textes que les données codées et structurées. Selon [Kahn, 1991], la représentation du temps est soit un flux continu ou un ensemble de valeurs discernables, soit une date ponctuelle *(moment)* ou un intervalle *(restreint par des limites temporelles supérieures et inférieures)*. La RI dans le DPI est souvent, pour ne pas dire presque toujours, liée à cette dimension (par exemple, *déterminez les différents épisodes d'une décompensation d'un diabète*"). La possibilité d'effectuer un raisonnement sur les données temporelles aide à la prise de décision clinique.

3.2.2.3 (Re)Formulation des requêtes utilisateurs

Pour être utile, un SRI doit être adapté aux utilisateurs finaux. Cela implique, d'une part que le fonctionnement du système soit intuitif, et d'autre part, que le système réponde aux requêtes de l'utilisateur. Ces contraintes, complexité des interfaces de RI et complexité des questions, ne sont pas les mêmes en fonction du cas d'usage [Terry et al., 2010].

3.2.3 Typologies des méthodes de RI

Ces outils de RI doivent prendre en compte l'hétérogénéité des données du DPI pour adopter des méthodes de RI adaptées à chaque type de données.

11. L'objectif est d'évaluer différents aspects des SRI [dans un contexte monolingue ou multilingue] comme les méthodes d'indexation du contenu textuel des documents médicaux.
12. Cross Language Evaluation Forum, URL : http ://clef-campaign.org/
13. I2B2, URL : https ://www.i2b2.org/NLP/Coreference/Main.php
14. Text REtrieval Conference (TREC), URL : http ://trec.nist.gov/

3.2.3.1 Vers une limite des données codées et structurées ....

Cette approche RI exploite uniquement les données codées et structurées pour répondre au besoin informationnel du soignant.
Le framework **I2B2** [Deshmukh et al., 2009] exploite les données codées, les données biologiques [et d'autres données de la génomique]. Le domaine d'application concerne la RI clinique (sélection de cohorte) et l'analyse statistique des données.
Le prototype **RIDoPI** (cf. section 5.3 du chapitre 5), à l'état actuel, exploite les données PMSI et la biologie pour rechercher des éléments du DPI.
Ces données sont stockées dans une base de données et ce type de RI est plus simple (recherche sur des éléments de données de type *attribut-valeur*). Les données sont représentées par un modèle d'information [de données], ce dernier doit être adapté à la RI. L'évaluation de [Koopman et al., 2011] a montré la limite de ces types de données (codées) car elle n'apporte pas de précision dans la RI. Certaines limites tiennent aux terminologies (*les codes CIM10 correspondant aux infections urinaires sont situés dans le chapitre des maladies de l'appareil génitourinaire mais absents du chapitre des maladies infectieuses.*). La requête de l'utilisateur sur les patients atteints de maladies infectieuses sera incomplète et va donc ramener des résultats incomplets [Lieberman, 2010].
Il est, donc, nécessaire de recourir à d'autres méthodes d'indexation (*a fortiori* de RI) pour exploiter le contenu textuel des documents.

3.2.3.2 ... à l'exploitation des données textes

A la différence de la première, cette seconde approche exploite le contenu textuel des documents du DPI. Elle peut se faire de manière différente : une RI *plein texte* comme [Cuggia et al., 2010, 2011] ou une RI fondée sur les métadonnées décrivant la sémantique du contenu textuel [Murphy et al., 2006]. Notre prototype **RIDoPI** se fonde sur la RI fondée sur les métadonnées. Ces métadonnées sont des termes d'indexation appartenant à une ou plusieurs terminologies de référence.
[Currie et al., 2001] proposent une approche linguistique (variation lexicale des termes, prise en compte du contexte) pour analyser les documents médicaux afin d'identifier les patients qui ont des problèmes de coeur et qui fument.
[Jain et al., 2010] proposent une méthode d'expansion d'une requête en utilisant plusieurs sources de connaissances, incluant les relations sémantiques fondées sur des ontologies, des méthodes d'apprentissage supervisé des co-occurences d'un terme.
[Plaza and Díaz, 2010] utilisent l'outil MetaMap pour exploiter les relations sémantiques d'UMLS afin de rechercher des cas similaires de DPI.
Ces méthodes TAL et sémantiques sont largement utilisées pour améliorer la pertinence des résultats par rapport aux méthodes lexicales [Plaza and Díaz, 2010].

[Hripcsak et al., 2009; Cuggia et al., 2010] ont montré l'apport de l'indexation *plein texte* pour ce type de RI. L'étude d' [Hripcsak et al., 2009] a démontré la possibilité d'utiliser le texte narratif pour la surveillance en milieu ambulatoire. Les travaux de [Cuggia et al., 2010, 2011] ont montré l'apport de la RI *plein texte* en plus des données PMSI.

3.2.3.2.1 Combinaison des deux méthodes de RI

De nombreuses études ont comparé les performances des indexations TAL et PMSI [Schulz et al., 2008; Cuggia et al., 2010, 2011; Murff et al., 2011], leurs résultats sont assez hétérogènes et ne permettent pas de conclure à la supériorité d'une stratégie par rapport à l'autre. Nous pouvons penser que l'amélioration des connaissances en TAL permettra d'améliorer les performances des outils d'indexation automatique et, subséquemment, des outils de recherche d'informations.
La combinaison des deux méthodes de RI apportent une plus-value en terme de précision [DeLisle et al., 2010; Cuggia et al., 2010, 2011]. Ces derniers combinent les données issues du codage PMSI et celles issues de l'indexation des données textuelles dans leur moteur de recherche.
Un cas particulier de RI [Farfan et al., 2009; Hristidis et al., 2010] qui ont développé des outils permettant de rechercher des informations dans un corpus de documents CDA HL7. Ces outils implémentent les techniques "'*Authority Flow*" pour exploiter la structure XML du document dans le processus d'indexation et de RI. Quant à [Liu et al., 2009], ils utilisent une approche ontologique pour rechercher dans des documents CDA HL7 et exploitent les relations sémantiques pour améliorer le rappel.

3.2.3.3 La prise en compte de l'utilisateur

Certains SRI se sont focalisés sur les requêtes utilisateurs et implémentent certaines approches de la RI orientée *Web* : des SRI type QR [Yu and Yilayavilli, 2009], des interfaces de requêtes paramétrables [Austin et al., 2008], l'exploration à travers un vocabulaire standard pour formuler une requête [Joubert et al., 1996; Cuggia et al., 2010], des *infosbuttons* pour générer la requête [Zeng and Cimino, 1997].
[Joubert et al., 1996] ont développé un système permettant de définir les requêtes à partir de l'UMLS afin d'interroger des données du DPI.
L'*infobutton* est utilisé par le soignant lors de la consultation d'un rapport de radiologie afin de formuler, à partir de templates génériques de requêtes, sa question (par exemple, *Est-ce que ce <disease/finding> apparait dans les autres rapports de radiologie du patient ?*) [Zeng and Cimino, 1997].
[Austin et al., 2008] ont proposé un modèle d'information générique de requêtes spécifiques au domaine du cancer afin d'implémenter des interfaces de requêtes paramétrables

pour répondre à ce type de requêtes : *"Trouver l'âge et le sexe des patients qui ont une maladie d'Hodgkin, pour lesquels le diagnostic initial a eu lieu entre 50 et 70 ans inclus"*. Un moteur de recherche de type QR (sur des métadonnées plutôt que sur des mots) a été développé par [Yu and Yilayavilli, 2009] pour rechercher dans les dossiers patients.

3.3 Comparaison des systèmes de RI existants

Comme nous l'avons exposé, plusieurs approches pour l'indexation, l'interrogation et l'analyse des données ont été implémentées dans les SRI. Le tableau 3.2 a pour objet de résumer les principales caractéristiques de quelques systèmes : **R-oogle** [Cuggia et al., 2010], **I2B2** [Murphy et al., 2006], **MIRS** [Spat, 2007] et **XOntoRank** [Farfan et al., 2009].
Les caractéristiques comparées sont :

- les types de données (données diagnostiques et d'actes [PMSI], données biologiques [BIO], courriers et comptes-rendus médicaux [CRH], médicaments [MEDI], le type de documents (document non structuré [texte], document semi-structuré [structuré], ...) ;
- les méthodes d'indexation (indexation conceptuelle et/ou sémantique [basé-concepts], indexation textuelle [basé-terme]) et les outils d'indexation (Lucene, F-MTI, ...) ;
- les terminologies et ontologies médicales utilisées ;
- le type de RI (rechercher *plein texte* [RI texte], rechercher sur des métadonnées des termes d'indexation et d'attributs [RI structurée]) ;
- autres traitements inhérents à cette RI.

3.3.1 R-oogle [Cuggia et al., 2010]

R-oogle est une plateforme française du CHU de Rennes dédié à la RI au sein de DPI. Cette plateforme est constituée d'un entrepôt de données stockant les données du DPI (données biologiques [LOINC], données d'actes [CCAM], des comptes-rendus médicaux de radiologie, d'anatomopathologie et des courriers de sortie) et d'un ensemble d'outils de RI combinant des méthodes de RI *sémantique* basées sur l'exploitation des métadonnées spécifiques aux documents (métadonnées issues du SIC, métadonnées sur la structure logique du document [15]), et des méthodes de RI *plein texte* exploitant le contenu textuel.
R-oogle comprend 3 modules :

- un module d'intégration des données du DPI dans l'entrepôt de données utilise l'open source *Talend* [16] ;
- un module de pré-traitement incluant une indexation *plein texte* fondée sur **Lucene** et des méthodes TAL avec l'extracteur **NOMINDEX** [Pouliquen, 2002]) effectuant une expansion *sémantique* de cette indexation, à l'aide du thésaurus MeSH et les synonymes français de l'UMLS [17] ;

15. Découpage en bloc du contenu textuel
16. Open Source Intégration Software and Data Management - Talend, URL : http ://fr.talend.com/index.php
17. les documents sont indexés par les termes qu'ils contiennent, mais aussi, lorsqu'un de ces termes

– un module de recherche multi critères permettant de faire : (1) une RI *plein texte* avec une gestion de l'aspect temporel (*une RI sur plusieurs documents pour même séjour*), et (2) une RI *structurée* à travers des interfaces permettant de naviguer dans des vocabulaires standards (LOINC, CCAM, MeSH) en fonction du type de données. Ce module offre aussi une représentation temporelle de l'intégration des données d'un DPI dans un diagramme de GANTT.

3.3.2 I2B2 [Deshmukh et al., 2009]

I2B2 est une plateforme open source, appelé "Informatics for Integrating Biology and the Bedside" (I2B2) [18] [Murphy et al., 2006], développée aux Etats-Unis et dédiée à la recherche translationnelle. Ce framework est implémenté dans plusieurs pays [Takai-Igarashi et al., 2011; Ganslandt et al., 2011] et pourrait devenir un *standard de facto*. **I2B2** stocke les données dans un entrepôt de données centré sur l'exploitation de données structurées, contrairement à **R-oogle** qui a montré l'apport de la recherche plein texte dans un entrepôt de données [Cuggia et al., 2010].
L'architecture fonctionnelle de cet outil, *orienté-service* (SOA) [19], propose une organisation en module (*hive*), chaque module ayant une fonction spécifique au sein de l'applicatif. Cette modularité facilite l'utilisation des données cliniques et génomiques des DPI pour les chercheurs. Ces modules incluent :

– des outils TAL pour l'indexation et l'extraction de concepts basés sur des ontologies médicales ;
– des outils de RI pour sélectionner des patients sur des données codées et structurées. L'aspect temporel est plus ou moins intégré ;
– des outils graphiques pour réaliser des analyses statistiques sur les résulats de la sélection.

3.3.3 MIRS [Spat, 2007]

MIRS (Medical Information Retrieval System) est un prototype d'interfaces de RI sur des documents médicaux non structurés de DPI de l'hôpital de Styria. Ces documents, anonymisés et rédigés en langue allemande, incluent des comptes-rendus de cardiologie, lettres de sortie, ... **MIRS** est un outil de RI combinant à la fois des méthodes de RI *pleine texte* et des méthodes de classification automatique de document. Il est composé de trois modules :

appartient à un vocabulaire contrôlé, par les synonymes et les pères de ce terme au sein de l'UMLS

18. http ://www.i2b2.org

19. Forme d'architecture de médiation ou de modèle d'interaction applicative mettant en œuvre des services (Service Orientied Architecture)

- un module de classification fondé sur une extension du framework **WEKA** [20] pour une classification supervisée à *multi-valeurs* des documents médicaux. Cette classification permet d'assigner pour chaque document, une ou plusieurs catégories (qui représentent généralement les spécialités médicales) desquelles ils relèvent : des métadonnées *prédictives*.
- un module de pré-traitement fondé sur les fonctionnalités de **Lucene** pour une indexation *plein texte* du contenu textuel des documents, étendu sémantiquement avec l'algorithme de classification. Le corpus d'index des documents inclus en plus des métadonnées issues de l'indexation (et des métadonnées spécifiques de la base de DPI) les métadonnées *prédictives* ;
- un module de recherche incluant : (1) des fonctionnalités d'interprétation et d'appariement de la requête utilisateur avec le corpus de documents, (2) des interfaces de RI pour la formulation de ces requêtes et le paramétrage de classification des résultats. Ces métadonnées *prédictives* sont exploitées par l'utilisateur, au moment de sa requête, pour favoriser le rappel des documents relevant d'une ou de plusieurs spécialités.

3.3.4 XOntoRank [Farfan et al., 2009]

La RI dans les documents XML consiste à identifier les éléments XML [21] (le noeud de l'arbre XML) les plus pertinents par rapport aux termes d'une requête utilisateur. La majorité des approches proposées dans la littérature sont des adaptations des modèles traditionnels (vectoriel, probabiliste, sémantique, ...). Nous présentons un de ses systèmes : **XOntoRank**(Ontology-Aware Search of Electronic Medical Records).
XOntoRank est un moteur de recherche permettant de faire une RI *sémantique* dans des documents médicaux structurés conformes au standard HL7 CDA. Ces documents ont la particularité de contenir à la fois des données codées, structurées et des données textes. Ce système comprend deux phases :

- une phase d'indexation *sémantique* et d'extraction des concepts dans le contenu textuel des documents CDA HL7 à l'aide de la nomenclature SNOMED. **XOntoRank** utilise un système de pondération des concepts d'indexation SNOMED implémentant deux algorithmes : (1) l'algorithme Breadth-First-Search [22] (BFS) pour calculer le poids (*p1*) du concept pour cet élément XML en se basant sur un principe de propagation de pertinence à travers le graphe des noeuds du document

20. Weka 3 - Data Mining with Open Source Machine Learning Software in Java, URL : http ://www.cs.waikato.ac.nz/ml/weka/
21. Un élément XML est considéré comme un document pour appliquer les fonctions de RI.
22. Breadth-First-Search, URL : http ://en.wikipedia.org/wiki/Breadth-first_search

XML [23] et (2) l'algorithme Okapi BM25 [Hersh, 2008] pour calculer le poids (*p2*) du concept dans le corpus des documents CDA HL7. Le résultat de cette phase d'indexation est une base de documents XML indexés avec, pour chaque élément XML, un concept SNOMED et un poids (*max(p1,p2)*) ;

– une phase de requête dans laquelle les concepts SNOMED des termes extraits de la requête utilisateur sont mappés avec la base de documents XML indexés. **XOntoRank** utilise l'algorithme **XRANK** [Guo et al., 2003] pour retourner la partie XML du document CDA HL7 correspondant à la requête.

XRANK, une variante de l'algorithme PageRank [24], pour exploiter la structure du graphe des noeuds du document XML dans le ré-ordonnancement de la liste des résultats. A la seule différence d'**XRANK**, [Hristidis et al., 2010] prennent en compte les liens entre les éléments XML des documents CDA HL7 [25].

23. Le score de pondération d'un noeud-père est calculé à partir des scores des éléments fils en utilisant une fonction d'agrégation

24. L'algorithme PageRank exploite les liens hypertextes pour mesurer la pertinence des pages, et donc améliorer la pertinence de la RI.

25. Ces attributs permettent de référencer des données au sein du même ou d'un autre document HL7 CDA

	R-oogle	**I2B2**	**MIRS**	**XOntoRank**
Type de données	PMSI, BIO, CRH	PMSI, BIO, MEDIC	PMSI, BIO, CRH	CDA HL7
Type de document	texte	texte	texte	structuré
Type d'indexation	basé-concept	basé-concept	basé-terme	basé-concept
Outils d'indexation	Lucene	-	Lucene	algorithme BFS
Vocabulaires	LOINC, CCAM, MeSH	Ontologies	-	SNOMED
Type RI	les deux	RI structurée	RI texte	les deux
Évaluation Score	-	-	-	Okapi BM25
Autres traitements	Visualisation du DPI (diagramme de GANTT)	Outils d'analyses statistiques	Classification Documents (**WEKA** étendu)	Classification Résultats (**XRANK**)

TABLE 3.2 – Évaluation de certains SRI dédiés aux dossiers patients informatisés

3.3.5 Avant-synthèse sur notre positionnement par rapport aux SRI de l'état de l'art

Ces outils présentés sont variés dans leur méthodologie de RI. Certains utilisent des techniques "TAL" ou de "data mining" fondées sur des métadonnées spécifiques ou standards pour enrichir la sémantique de leur indexation des documents (textes ou semi structurées).
Notre objectif de RI est de pouvoir faire une RI simple sur un dossier patient ou sur un ensemble de dossiers patients. Notre positionnement est d'être *compliant* avec l'outil **I2B2** dans ses fonctionnalités de RI en excluant les fonctionnalités d'analyses statistiques.. Une comparaison plus approfondie sera faite entre les outils **RIDoPI** et **I2B2** dans le paragraphe 6.2.2.3 du chapitre 6. Par ailleurs, nous proposerons un modèle de données générique spécifiquement adapté à la RI dans le DPI et pouvant intégrer le SI CISMeF et les données du PTS. Ce modèle devra nous permettre, en plus des données PMSI et données biologiques, d'intégrer les données textuelles sous forme de métadonnées.

Synthèse

Nous avons présenté, les principes et les approches, en général, des systèmes de recherche d'information existants, et décrit l'intérêt de la RI dans le dossier médical. L'intérêt principal est l'amélioration de la qualité des soins par l'amélioration des outils de travail des professionnels : un accès facile à l'information recherchée, spécifique au patient durant sa prise en charge ou l'accès à des informations cliniques sur un ensemble de patients pour la sélection de patients dans le cadre de la constitution de cohortes.
La première difficulté est l'expression du besoin informationnel toujours incomplète et peu précise des professionnels. Les questions cliniques que se posent les professionnels sont de deux types et indissociables : questions dont la réponse se trouvent dans les connaissances du domaine (ou de la spécialité) spécifique au patient pour [re]formuler et, questions dont la réponse se trouve dans le dossier patient informatisé.
Les outils de RI sont la seule réponse à ces dernières. Devant ce constat et face à l'hétérogénéité et au volume des données, se trouvant dans le dossier médical, un modèle de données adapté à la RI est nécessaire. Dans le chapitre suivant, nous aborderons la problématique de modèles spécifiquement adaptés à cette fonction.

Deuxième partie

Mise en oeuvre d'un modèle de données générique adapté à la RI au sein du DPI

CHAPITRE 4

Description d'un modèle de données générique

Introduction

Alors que les méthodes de modélisation des données sont dominées par les concepteurs, les terminologies sont développées par les cliniciens. Dans le chapitre 2, nous avons présenté les problématiques de la modélisation du DPI et nous avons montré que les modèles d'information [de données] qui définissent *la structure* de l'information à stocker dans un DPI et les modèles terminologiques qui définissent *le sens* de ce qui sera stockée, doivent idéalement être en mesure de travailler ensemble pour améliorer la qualité des données (cohérence, exhaustivité, ...), et donc d'accroitre leur réutilisation dans d'autres domaines, parmi lesquelles la RI [Rector et al., 2001].
Développer des fonctions de RI dans le dossier patient nécessite, outre les outils nécessaires à la RI et des données structurées ou codées, un modèle de données adapté voire dédié. Dans le chapitre 3, nous avons montré l'intérêt de la RI au sein du DPI.
La modélisation des données du dossier patient est [et sera] un exercice difficile du fait de l'étendue du domaine, de la complexité des connaissances et de la multidisciplinarité des acteurs (médecins, cliniciens, informaticiens, etc.) qui ont des approches différentes sur le processus de soin [Johnson, 1996]. *D'où la question - Comment pourrait-on faire pour qu'une information soit représentée sous une forme adaptée pour la RI clinique* [Los et al., 2005] ?
Dans ce chapitre, nous étudierons le modèle de données de CDP, la modélisation des données et les principaux modèles de DPI, nous décrirons notre démarche et le mo-

dèle auquel elle a abouti, enfin nous comparerons ce modèle aux modèles existants (voir Figure 4.1).

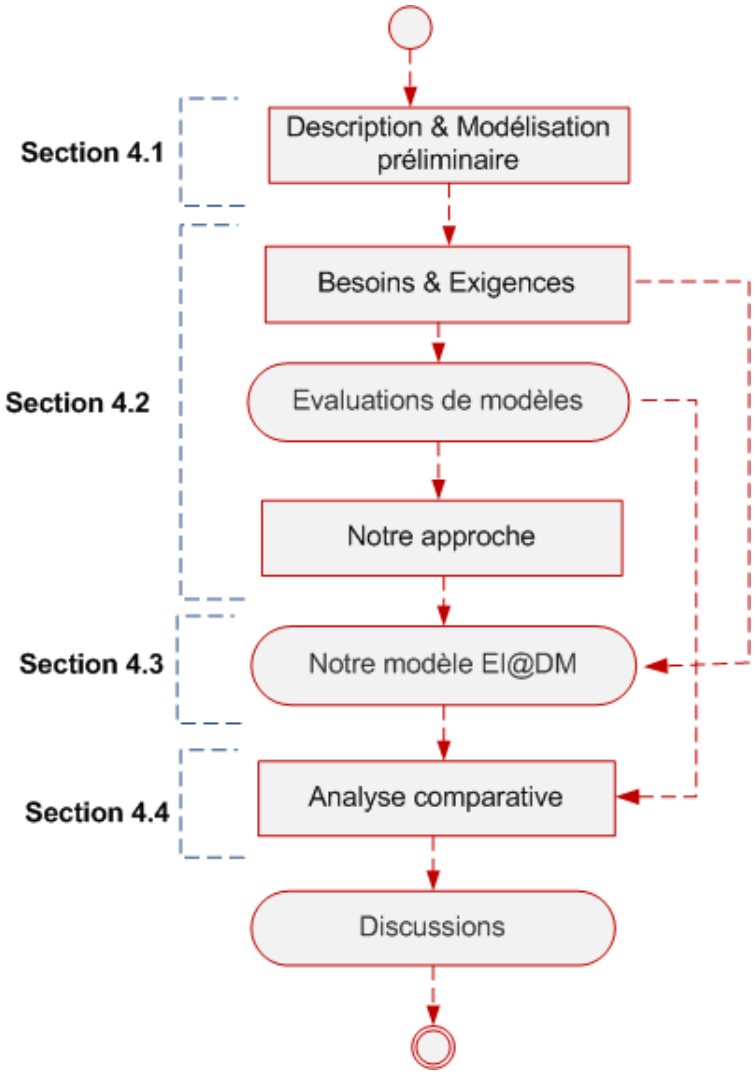

FIGURE 4.1 – Schéma synoptique de notre modélisation

4.1 Le DPI au CHU de Rouen

4.1.1 Informatisation du dossier médical

L'informatisation du dossier médical au CHU de Rouen a débuté en 1992 par la mise en place d'une application développée localement fonctionnant sur un site central [Massari and Fuss, 2000]. Elle est remplacée en 1999 par C-PAGE Dossier Patient (CDP), progiciel développé en partenariat avec le CESIH de Bourgogne et répondant à la norme européenne HISA. CDP fonctionne en mode client serveur sur une architecture 3-tiers se basant sur une base de données relationnelle, stockée physiquement sur Oracle. Il fonctionne avec trois modules principaux :

- un module clinique présentant les dossiers sous forme d'arborescence événementielle (patient, épisodes, séjours, actes) dont la présentation peut être adaptée selon le besoin de l'utilisateur par l'intermédiaire de vues ;
- un module biologique pour la gestion des examens biologiques ;
- un module de recherche de données fonctionnant sur la base de requêtes paramétrables.

4.1.2 Le DPI sous CDP

Le dossier patient est constitué d'éléments informationnels recueillis au cours des prises en charge hospitalières au CHU de Rouen. Un épisode de soins correspond au regroupement de 1 à n prises en charge qui peuvent être des hospitalisations ou à des périodes pendant lesquelles le patient a bénéficié de soins externes (concept de **Contact** du modèle HISA). Ces dossiers contiennent une masse importante de données et d'informations de différentes natures et de différents formats. Sont contenues dans ces dossiers les éléments suivants :

- au **niveau patient**, les données d'identification (nom, prénom, sexe, date de naissance, etc.), les données hospitalières (identifiant permanent, médecin traitant, etc.) ;
- au **niveau prise en charge**, les périodes de consultation et d'hospitalisation auxquelles sont rattachées les actes médicaux diagnostiques ou thérapeutiques et certains actes paramédicaux. Ils sont codés en CCAM depuis 2005, en CDAM auparavant. Les diagnostics de séjour sont codés depuis 1992 en CIM-9 puis en CIM-10 (depuis 1996) ;
- les résultats des examens biologiques comprenant les résultats, les bornes de normalité, ... Aucune terminologie internationale n'est encore utilisée mais la terminologie LOINC sera utilisée comme terminologie de référence à court terme ;

– certaines données de prévention et des données multimédia issues du PACS.

Ces données sont stockées sous forme structurée et codée pour certaines (données patients, données épisodes de soins, données biologiques, etc.) mais la grande partie de l'histoire médicale des patients est contenue dans les comptes-rendus médicaux et courriers stockés physiquement au format texte. L'application gère actuellement plus de 4 millions de comptes-rendus et autres courriers pour 800.000 patients différents.

4.1.3 Exemple d'un dossier patient sous CDP

La figure 4.2 représente un DPI sous CDP.

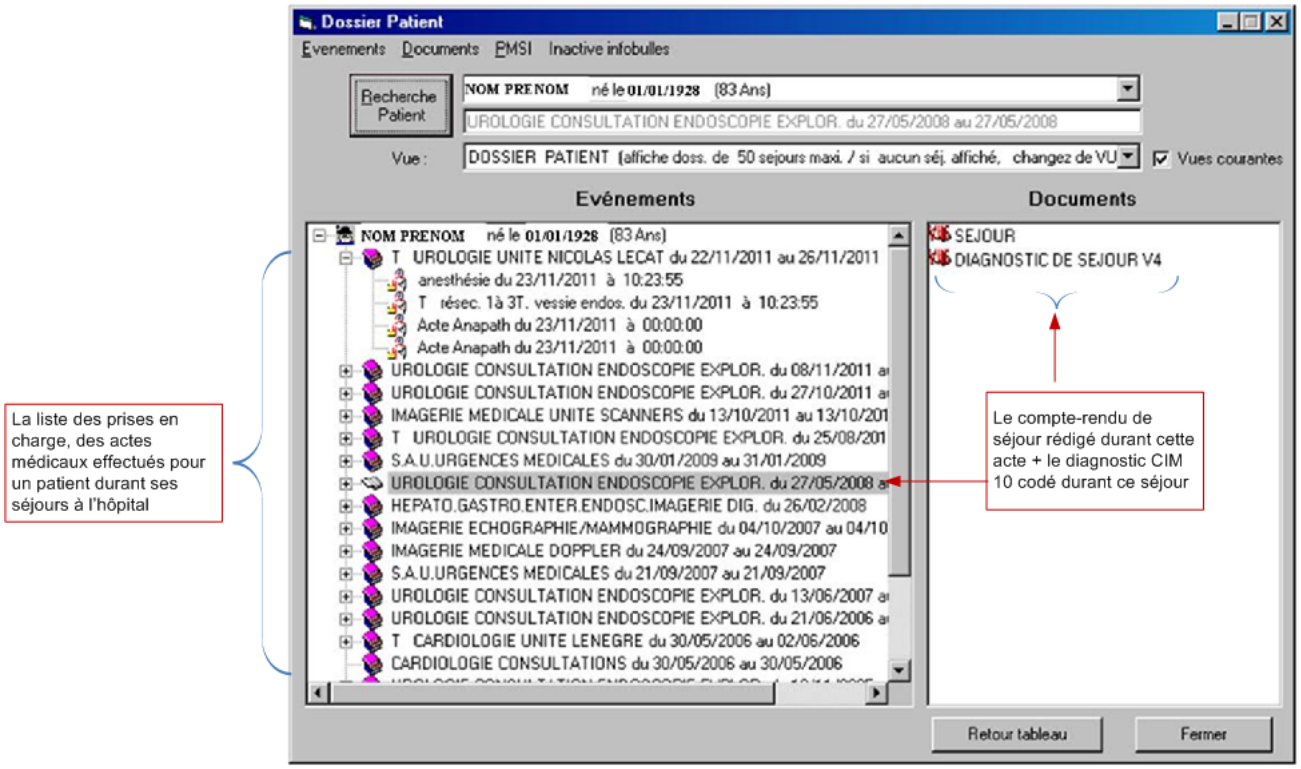

FIGURE 4.2 – Un dossier patient informatisé sous CDP

4.1.4 Modèle CDP réduit

La base relationnelle du système CDP contient plusieurs dizaines de tables gérant les données du dossier patient (100 à 110 tables) mais aussi les données nécessaires au fonctionnement de l'application CDP. La plupart des tables de CDP ne sont pas nécessaires à la RI, ce qui nous a amené à créer un sous-modèle regroupant dans un nombre minimal de table, les données pertinentes pour répondre aux besoins informationnels du médecin. Ce sous-modèle prend en compte les données patients, prises en charges, actes médicaux, courriers et comptes-rendus médicaux et les examens biologiques. Ce travail a été fait en collaboration avec l'expert du domaine, le Dr Philippe Massari (PM).

4.1.4.1 Description conceptuelle

Le sous-modèle du modèle physique de CDP est constitué d'un ensemble de 7 tables schématisées conceptuellement dans la figure 4.3. Le modèle conceptuel contient les entités suivantes :

- **Patient** permet l'enregistrement des données démographiques du patient (à l'exclusion des noms et prénoms du patient) ;
- **Séjour** gère les données concernant les prises en charge ;
- **Acte** gère les données concernant les actes médicaux ;
- **CodesDiagnostics** contient les codes CIM 10 des pathologies principales et associées diagnostiquées durant LE séjour du patient ;
- **CodesActivités** contient les libellés d'activités et les codes CCAM réalisés pendant LE séjour du patient ;
- **CodesAnalyses** permet l'enregistrement des résultats de laboratoires provenant des SGL et le code LOINC associé ;
- **Courriers** permet de gérer l'ensemble des textes, quel que soit leur niveau de rattachement.

4.1.4.2 Limites du modèle

Le modèle de données de CDP est un modèle relationnel "classique" avec de nombreuses tables, exploiter les données d'un ou de plusieurs dossiers ou y rechercher des informations nécessitent :

1. de connaître le modèle ;
2. de construire des requêtes spécifiques incluant de nombreuses jointures.

Ceci ne peut être fait par les utilisateurs sans l'aide d'un informaticien. Actuellement, au CHU de Rouen, seuls les praticiens formés sont en mesure d'interroger les données

structurées disponibles, en particulier les codes diagnostiques CIM-10 et les codes CCAM. Les interfaces de recherche consistent en des zones de saisie des critères de recherche sous forme de texte, qui permettent de rechercher des patients, des prises en charge et/ou des actes médicaux. Le modèle n'est pas adapté à des RI plus complexes (prise en compte des contraintes temporelles relatives) sur l'ensemble des données codées. C'est un modèle contraint par les exigences organisationnelles, techniques et fonctionnelles du SIC implémenté et donc inadapté à la RI utilisant des outils du Web Sémantique.

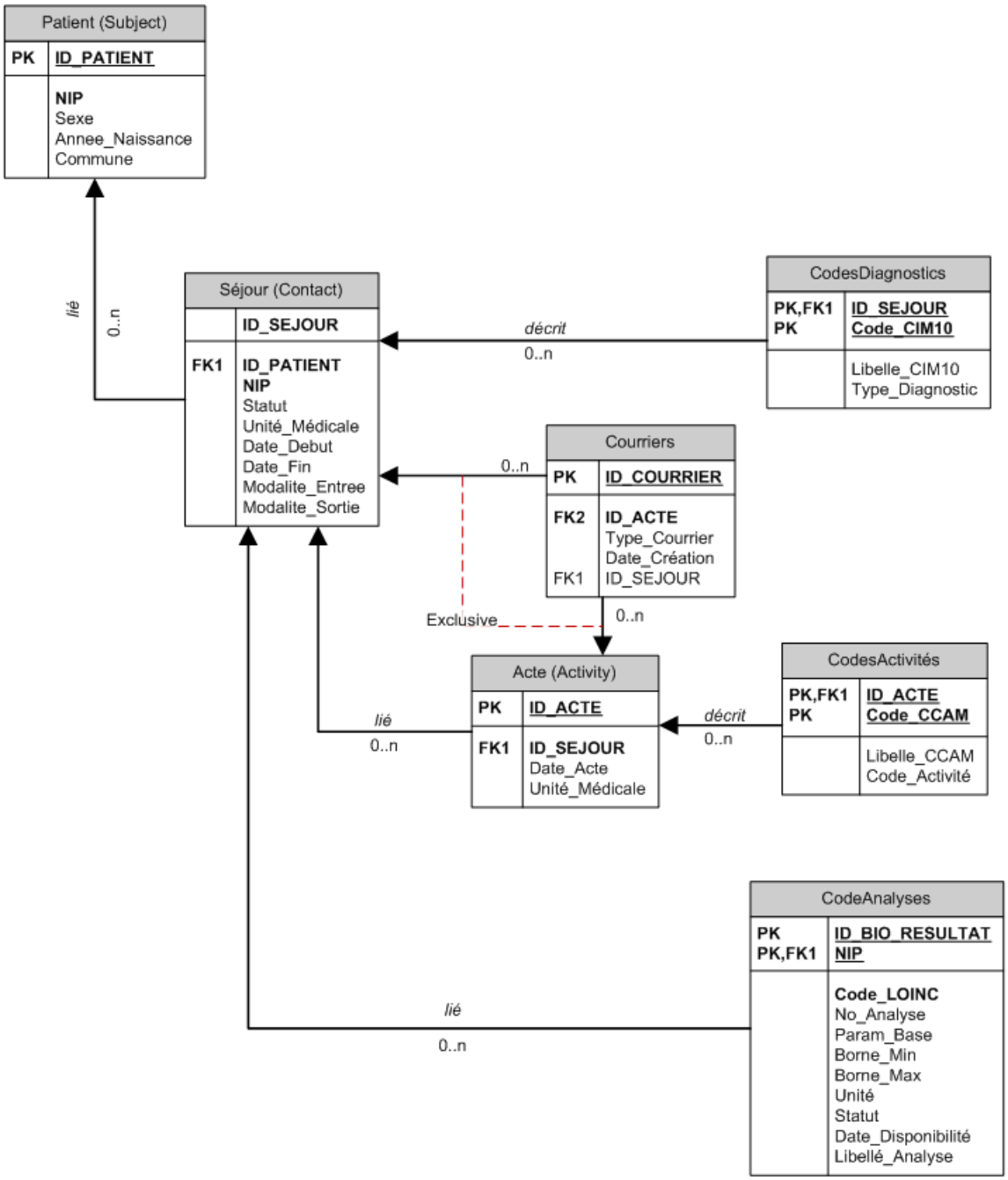

FIGURE 4.3 – Sous modèle de CDP

4.2 Paradigme de modélisation

Dans cette section, nous définirons les exigences et besoins pour notre modélisation, puis nous donnerons une évaluation succincte de quelques modèles et enfin, nous présenterons notre approche de modélisation.

4.2.1 Besoins et exigences

Le domaine médical est caractérisé par un certain nombre de particularités :

- **richesse des données et des connaissances** : la notion de *'concept'* inclus tout élément de l'informationnel au processus, le dossier patient permet d'avoir des informations précises et granulaires sur l'état de santé du patient ;
- **longévité de l'information** : les données médicales ont, pour la majorité, la même durée de vie que le patient qu'elles décrivent et au délà pour être [re]utilisées dans les recherches épidémiologiques ou à des fins médico-légales ;
- **besoin d'une interopérabilité élevée des systèmes** : l'information doit être partagée ;
- **besoin d'outils de traitements intelligents** : les données doivent être exploitables au niveau sémantique des structures de connaissances, pour les moteurs de recherches, les outils d'aide à la décision, etc.

La modélisation doit prendre en compte ces spécificités et faire face à la complexité, la variabilité du domaine médical en raison de l'amélioration des connaissances (ajout de nouveaux modèles de connaissances, ...), être adaptée aux processus cliniques et à l'émergence du DMP, ceci nécessite la prise en compte de certaines contraintes :

- **des contraintes sémantiques**

  (1) prévoir un niveau d'abstraction suffisant pour prendre en compte les niveaux des données, les details de la visite classique patient-médecin, les étapes des examens physiques durant la prise en charge ou les diagnostics, les procédures ou d'autres types et structures de données cliniqes mélangées à des concepts tels que des éléments calculés, agrégés par rapport à d'autres variables unitaires (par exemple le score d'APGAR), et donc éviter de perdre trop d'information ;

  (2) prévoir l'évolution du domaine ou l'intégration d'un nouveau domaine (par exemple, l'hospitalisation à domicile) en séparant nettement le niveau *"information"* et le niveau *"connaissance"* ; en effet, quand les exigences du domaine de connaissances changent, uniquement le modèle de connaissance change et non, le modèle de données ;
- **des contraintes techniques**

  (3) prévoir la facilité d'utilisation du modèle ; il est prévu que ce dit modèle

doive intégrer l'ensemble des données de manière intelligente afin de faciliter la RI, mais le but n'est pas nécessairement de satisfaire ce besoin. Ce modèle pourra être la base de gestion des données patients (en évitant d'accéder à la base CDP d'origine) et être support à l'exécution de requêtes de scénarios cliniques ou à d'autres outils, comme les outils d'aide à la décision, ... La base CDP ne sert qu'à la consultation des dossiers patients et n'est pas adaptée pour l'intégration de ces outils répondant aux besoins des professionnels : l'inclusion d'un patient dans un protocole d'essai clinique sera plus efficace sur une base contenant des données validées, codées et intelligemment structurées indépendamment du SI d'origine ; la gestion de la confidentialité des DPI n'est plus une question ; il est difficile de mettre à jour la base CDP composée de centaines de tables pour les besoins de RI (par exemple, la réindexation des données médicales) ; des analyses multidimensionnelles peuvent être exécutées sur cette base pour la RI épidémiologique.

(4) la structure de données de notre modèle doit présenter les informations dans un minimum de contraintes et de manière flexible pour faciliter sa maintenance.

Un modèle satisfaisant ces contraintes (ou critères) et intégrant *"la structure et la sémantique globale de l'ensemble des données"* doit être un modèle générique, flexible avec une expressivité suffisante pour représenter des données médicales complexes. Ce doit être un modèle multi-niveau afin de réaliser l'interopérabilité, non seulement au niveau des données mais aussi au niveau sémantique. Le modèle de données que nous proposerons répondra en premier lieu à cette définition et devra satisfaire ces contraintes.

4.2.2 Évaluation de l'adaptation à la RI des modèles existants

Dans cette section, nous détaillerons des modèles présentés dans la section 2.2.1 du chapitre 2, selon deux aspects : un point de vue *"médical"* pour connaître l'élément de base du modèle et un point de vue *"technique"* pour évaluer son adaptation à la RI, la complexité et la flexibilité du modèle pour une réutilisation. Nous avons restreint notre étude aux cinq principaux modèles (ou techniques) utilisés dans les DPI en séparant les modèles normatifs et les modèles non normatifs.

4.2.2.1 Modèles non normatifs

Ces modèles non normatifs utilisent plusieurs techniques différentes pour représenter la structure et la sémantique des données du DPI.

4.2.2.1.1 Le modèle PEN&PAD

[Rector et al., 1993; Goble et al., 1994] partent du principe que les informations contenues dans le DPI ne portent pas sur ce qui est *"vrai ou faux"* chez le patient mais sur ce qui a été observé et pris en considération par les soignants. Partant de ce principe, dans le cadre du projet PEN&PAD,Alan Rector et ses co-auteurs présentent un modèle pour gérer de manière formelle le DPI. Dans ce modèle, les auteurs ne font aucune distinction entre les symptômes, les signes, les analyses biologiques ou les traitements médicamenteux et considèrent le DPI comme un ensemble d'observations et de "meta observations". Ces dernières sont déduites des observations initiales par raisonnement et sont utilisés pour deux choses : enregistrer les décisions prises (par exemple, lier l'observation à un problème) et les dialogues cliniques (l'interaction entre les acteurs intervenant dans le processus de soin).
L'élément de base du modèle est *"une observation"* qui correspond à un ensemble d'états médicaux. Toute observation est liée à un `"objet observable"` qui peut correspondre à une radiographie, un patient, un échantillon de sang, ... Chaque état médical est sous une forme "Sujet-Attribut-Value" et chaque état correspond à une nouvelle instance dans le modèle. Tous les états médicaux se sont déroulés à un moment, en un lieu et avec un agent (professionnel ou appareil médical). Pour comprendre ce modèle, prenons comme exemple l'observation (E1) "*La fracture de Mr Jones observée était localisée au niveau du fémur. Elle est sévère et "spiroïde" et a été diagnostiquée il y a 3 heures dans le service d'Urgence*". Le formalisme *Structured Meta Knowledge* (SMK) [Nowlan, 1993] est utilisé dans ce modèle. SMK est un système *compositionnel* pour décrire des concepts médicaux basés sur une description structurée sous forme de réseaux sémantiques. Ce réseau consiste en des ensembles :

- de noeuds appelés *les entités*, pour (E1) : "Fracture du fémur" (N1), "Mrs Jones" (N2) "Service Urgence" (N3) sont trois noeuds ;
- et d'arcs appelés *les relations*, pour (E1) : "de" (A1) et "diagnostiqué" (A2) sont deux arcs pour donner N1-A1-N2 [1] et (N1-A1-N2)-A2-N3 [2]. Thèse de [Nowlan, 1993]

SMK permet de modéliser le dossier patient en 3 niveaux d'abstraction :

1. le niveau **Categories** pour représenter la sémantique du dossier patient à l'aide de bases de connaissances. Pour (E1), on a trois concepts : Patient (C1), Fracture du Fémur (C2) et Service urgence (C3) ;
2. le niveau **Individuals** pour représenter les instances des concepts sur lesquelles les observations sont faites (patient, lieu, examen biologique, ...) et il permet d'obtenir une vue synthètique du DPI. Pour (E1), on a une instance pour chacun de ces concepts : Mr Jones (C1), la fracture du fémur de Mr Jones (C2), la fracture du fémur de Mr Jones observé dans le service d'Urgence (C3) ;
3. le niveau **Occurences** pour représenter les différentes occurrences des observations. Pour (E1), on a UNE occurence pour ces trois concepts qui correspond à la fracture du Fémur de Mr Jones observée il y a 3 heures dans le service d'urgence (C1,C2,C3) = E1.

D'autres niveaux sont définis pour permettre au modèle d'être intégré dans un environnement distribué. Le modèle est un modèle flexible pour enregistrer l'action clinique, la réflexion et le dialogue. Il propose seulement une ontologie d'informations limitée notamment sur les observations directes et les "meta-observations" pour faire la distinction entre les faits et les opinions.
Ce modèle est trop complexe pour être implémenté dans notre SI CISMeF. Nous n'avons pas choisi d'adopter une méthodologie formelle pour modéliser le contenu de notre DPI. Nous avons donc considéré que ce modèle n'était pas adapté à notre problématique.

1. Fracture du fémur de Mrs Jones
2. Fracture du fémur de Mrs Jones, diagnostiquée dans le service Urgence

4.2.2.1.2 Le modèle CLEF Chronicle

Dans le cadre du projet CLEF, [Rogers et al., 2006] ont développé un modèle formel pour avoir une représentation explicite des données cliniques, appelé *"Patient Chronicle Model"*. L'objectif de ce modèle est de pouvoir exécuter des requêtes complexes sur les données agrégées du patient pour la recherche clinique, d'offrir un résumé de l'histoire médicale du patient et de résoudre les problèmes de *'co-références'*[3] parmi les documents cliniques. Ce modèle est implémenté comme un réseau sémantique de données, compatible avec un modèle orienté objet comprenant deux parties : une partie générique pour la représentation temporelle des données, selon une approche SNAP/SPAN [Grenon and Smith, 2004] et les concepts issues des sources externes de connaissances autour d'ontologies, et une autre partie spécifique dérivée de ces ontologies. Dans cette approche SNAP/SPAN :

- **un concept SNAP** correspond à une prise en charge, une opération chirurgicale, une radiographie donnée ou un score : *des événements réalisés à un moment donné* ;
- **un concept SPAN** correspond à *un événement qui s'est répété plusieurs fois dans l'évolution du patient*, par exemple plusieurs consultations pour la même pathologie. Une tumeur ou une pathologie A sont des événements **SPAN**. Pour la tumeur, le **SPAN** est composé d'un ensemble d'événements **SNAP** correspondant à la même tumeur à des moments précis dans le temps ; pour la pathologie A, nous aurons un ensemble de consultations dans lesquelles nous pouvons voir évoluer l'état de cette pathologie. Par conséquent, un concept **SPAN** peut agréger les différentes valeurs des événements **SNAP** (pour la tumeur, l'évolution de sa taille, dans cet intervalle de temps, peut être calculée).

L'objectif de cette modélisation est d'avoir une instance UNIQUE d'un événement dans l'historique de santé du patient, car ces auteurs considèrent le DPI comme un ensemble d'événements **SPAN**. Reprenons l'exemple de l'observation (E1), nous aurons un événement **SNAP** pour le diagnostic de la fracture du fémur gauche de Mr Jones à la date D1. Si cette observation se répète à la date D2 (la même frature), le diagnostic de la fracture du fémur gauche de Mr Jones devient un événement **SPAN**. Par conséquent, nous aurons une seule OCCURENCE pour le concept "*Fracture du fémur gauche*", celle de Mr Jones enregistrée à la date D1 et D2.
Dans un scénario de RI, on cherchera les "Chronicles" associés au terme "Fracture du fémur gauche" et on est sûr qu'on aura une SEULE occurence d'un événement **SPAN**

3. Des heuristiques sont appliquées sur les types de données du modèle pour conserver l'unicité d'un événement clinique. Ces heuristiques permettent aussi d'éviter de fusionner des événements cliniques par exemple *"deux grosesses"* séparées de plus de 10 mois

pour ce terme observé aux dates D1 et D2.
Ce modèle est dédié à la recherche clinique et offre une vue oriented-time du DPI améliorée selon l'approche SNAP/SPAN.
Certes "*Patient Chronicle Model*" est un modèle générique avec une forte expressivité, pouvant servir comme support à la visualisation du DPI, à la synthèse automatique du dossier patient. Cependant, dans l'état actuel, par rapport à la disponibilité des données de la base CDP et de la technicité de nos outils d'indexation, il n'est pas facile d'implémentation et non flexible d'un point de vue technique.

4.2.2.1.3 Le modèle RDF

[Lindemann et al., 2009] ont utilisé le modèle RDF pour représenter l'ensemble des données du DPI. Le stockage des enregistrements est réalisé de telle manière qu'aucune information ne soit perdue. Cependant, ce modèle n'est pas adapté pour un grand volume de données (1 ligne du modèle relationnel est transformée en plusieurs triplets donc plusieurs lignes dans le modèle RDF). Chercher sur une colonne d'un attribut donné est susceptible de prendre une quantité considérable de temps en raison du nombre relativement important de lignes dans le modèle RDF. Par ailleurs, ce modèle est contraint techniquement à utiliser le langage SPARQL [4], le plus adapté pour le moment, pour interroger les données de triplets et exécuter des requêtes complexes. Le modèle RDF est analogue à une méthode de modélisation relationnelle. Les technologies du Web Sémantique constituent des réponses à certaines limites du modèle relationnel afin de mieux gérer les données structurées. Elles offrent un environnement pour manipuler des données et des informations et effectuer des inférences sur celles-ci. On peut les utiliser dans un modèle relationnel (notamment grâce aux opérateurs sémantiques d'Oracle) et donc, nous avons choisi de modéliser nos terminologies avec le modèle RDF pour utiliser une RI sémantique sur le DPI.

4. D'autres langages existes, URL : http ://www.w3.org/2001/11/13-RDF-Query-Rules/

4.2.2.1.4 Le modèle I2B2

Nous avons présenté, dans le chapitre 2, plusieurs modèles orientés DW pour représenter le DPI à des fins d'analyses. Nous allons particulièrement évaluer le modèle I2B2 [Murphy et al., 2006]. Ce modèle permet d'intégrer des données médicales diverses (biologie, génétique, clinique, ...) afin de pouvoir faire des interrogations multicritères dans le temps. L'architecture fonctionnelle de cet outil, orientée services (SOA), propose une organisation en cellules, chaque cellule ayant une fonction spécifique au sein de l'applicatif. Le modèle I2B2 est stocké dans le module *Clinical Repository Cell (CRC)*[5] correspondant à l'entrepôt de données qui stocke les données cliniques. Ce modèle est construit sur l'approches Entity-Attribute-Value (EAV)[6].
Dans ce type de modèle (DW), le concept le plus important est d'identifier ce qui constitue *un fait*. Pour I2B2, un fait est une observation sur un patient et la table de faits contient toutes les observations réalisées durant le suivi d'un patient. Les tables dimensions sont : la dimension Patient, la dimension Provider (professionnels), la dimension Visit (prise en charge) et la dernière : la dimension Concept (Ontologie).
Cependant, dans I2B2, les ressources ontologiques et les données médicales sont gérées indépendamment, par un lien bidirectionnel qui rend l'évolutivité de l'outil complexe. [Mate et al., 2011] proposent une approche ontologique pour intégrer des données hétérogènes dans ce modèle.
I2B2 n'est pas un modèle multidimensionnel standardisé, ni un système d'interrogation prenant en charge les relations dans les ontologies (subsomption, synonymie, ...). De plus, la modélisation multidimensionnelle proposée est fortement orientée sur le couple *patient-acte*, ce qui réduit le nombre de mesures associé aux actes médicaux utilisables [Choquet et al., 2008].
Le modèle I2B2 est limité dans la représentation d'une observation clinique durant une visite en un ensemble de faits (par exemple, représenter le score d'APGAR). Plusieurs faits correspondant au calcul du score d'APGAR seront enregistrés dans la table de faits mais les liens entre ces faits ne seront pas stockés [Deshmukh et al., 2009].
Nous n'avons pas utilisé ce modèle ainsi que les modèles similaires à ce dernier : OMOP Common Data, HMORN Virtual DataWarehouse, HIMSS Data Model. Néanmoins nous allons mettre en œuvre une approche de modélisation semblable à l'approche EAV utilisée dans I2B2 pour stocker nos métadonnées. Une comparaison de conformité de nos outils de RI sera réalisée avec ceux d'I2B2 (cf. paragraphe 6.2.2.3 du chapitre 6).

5. Voir la sous section 3.3.2 pour le détail technique du framework.
6. http ://en.wikipedia.org/wiki/Entity-attribute-value_model

4.2.2.2 Modèles normatifs

Ces modèles normatifs sont basés sur ce qu'on appelle des "Reference Model (RM)". Par définition, un RM est une représentation formelle spécifique du domaine avec un niveau élevé d'abstraction pour couvrir l'ensemble des concepts, des acteurs et des informations. Dans le domaine médical, un RM ne contient pas seulement les éléments génériques du domaine (structure et type de données, les modèles démographiques, les workflows) mais aussi les éléments spécifiques au DPI.
Ces RM sont des modèles d'information génériques *orientés communication* de dossiers patients informatisés entre les SIS. Nous présentons, dans ce paragraphe les modèles d'information `RIM` et `HISA` ensemble et le RM `OpenEHR` à part.

4.2.2.2.1 Modèles RIM&HISA

La norme EN 14822 connue sous l'acronyme RIM (Reference Information Model), version V3 de la norme HL7, est le modèle de référence pour définir les classes et attributs qui représentent les données cliniques et leurs relations. Ce RM permet de décrire de manière formelle le contenu des messages HL7 pour les différentes parties voulant interopérer et donc échanger des données sous forme de messages HL7. Les messages HL7 s'appuient sur un sous-modèle qui dérive du RM appelé Domain Message Information Model(DMIM)) [7]et contient les classes nécessaires pour spécifier ces messages.
Le RIM a fait l'objet d'une normalisation ISO 21731 :2006 et est présenté comme une référence qui peut faciliter l'harmonisation des standards d'informatique de santé. Il est fondé sur 6 classes principales constituant l'ossature du RM :

- Acte (**Act**) représente les actions (acte clinique et ce qui leur est lié) qui sont prévue, en cours, ou qui ont été exécutées et qui doivent être documentées comme actes de soins ;
- Participation (**Participation**) exprime le contexte dans lequel l'acte a été réalisé par qui, pour qui, le lieu de réalisation ;
- Entité (**Entity**) représente des personnes physiques ou morales (patient, médecin, institution, etc.) ;
- Rôle (**Role**) définit les rôles que les entités jouent comme participants à l'acte de soins ;
- Relation entre Actes (**ActRelationship**) ;
- Les liens entre différents rôles (**RoleLink**).

La prénorme européenne prENV 12967 connu sous l'acronyme HISA (Health Informatics Service Architecture) est un composant fonctionnel, un *"middleware santé"*, pour

7. http ://wiki.hl7.org/index.php ?title=Domain_Message_Information_Model

structurer le développement d'applications au sein des SIS. HISA a défini un ensemble d'informations spécifiques de santé et de services fondamentaux communs "Healthcare Common Services" (HCS) pour tout SIS. Cette prénorme se base sur un RM pour représenter de manière formelle chaque HCS. Les nombreuses classes de HISA RM sont regroupées en six HCS principaux services. Pour chaque HCS, un modèle d'information définit les classes, les attributs et leurs relations entre eux et avec les autres classes des autres groupes. Les six HCS sont :

- (**S**ubject-**HCS**) ;
- (**H**ealth **C**haracteristic-**HCS**) ;
- (**A**ctivity-**HCS**) ;
- (**R**esource-**HCS**) ;
- (**A**uthorization **H**ealth-**HCS**) ;
- (**C**oncepts-**HCS**).

Cette figure 4.4 représente la convergence de ces deux modèles.

Ces modèles permettent de définir, à un niveau élevé d'abstraction, le contenu des

HISA RM	RIM RM	Description
Activity-HCS	Act	Activités
Activity-HCS Health Characteristic-HCS	Act ActRelationship	Continuité des soins
Activity-HCS Concepts-HCS	Act Entity	Connaissances cliniques
Subject-HCS Resource-HCS Authorization Health-HCS	Participation RoleLink	Responsabilités dans la continuité des soins
Concepts-HCS	Entity	Gestion des données de santé

FIGURE 4.4 – Exemples de métadonnées

données échangées. Par exemple, pour ces modèles, l'Acte (Act pour RIM et Activity pour HISA) recouvre toutes les actions ou événements d'un service de soins tels que l'acte clinique ou la transaction financière. Le recours à la spécialisation intervient uniquement lorsque les concepts souhaités doivent avoir des attributs supplémentaires que les classes principales n'ont pas, ou des relations qui n'existent pas au niveau de ces classes.

Bien que détaillés pour le domaine clinique, ils présentent un certain écart avec les bonnes pratiques de modélisation, en spécialisant trop les concepts. [Smith and Ceusters, 2006] les considèrent complexes et ils ne sont pas appropriés pour une modélisation conceptuelle

(ici pour le DPI).
Ils ne nous apparaissent pas adaptés à la RI sur le DPI, c'est donc pour cette raison que nous n'avons pas utilisé ces deux modèles.

4.2.2.2.2 Modèle OpenEHR

La fondation OpenEHR a été créée pour le développement de spécifications standards et ouvertes, d'application et de sources de connaissances pour les SIS et en particulier pour le DPI. L'originalité d'OpenEHR est d'avoir introduit la notion d'archétype [Beale, 2002]. Cette nouvelle approche dissocie le modèle d'information en deux niveaux : le RM et les archétypes proprement dits. OpenEHR fournit un ensemble de modèles d'information (OpenEHR RM l'équivalent du RIM HL7 et HISA RM) et une base terminologique fondée sur des archétypes (modèle formel réutilisable et extensible) [8]. Le modèle d'information pour le DPI est organisé autour de 6 classes fondamentales [Beale et al., 2007] :

- La classe **EHR** est le concept le plus haut contenant l'ensemble du contenu du DPI ;
- La classe **TRANSACTION** : élément informationnel de base du RM ; il peut correspondre à un document CDA de la norme HL7. Une transaction peut être une prise en charge, un compte-rendu de radiologie, un examen biologique, ... ;
- La classe dossier (**FOLDER**) permet de classifier les transactions dans une hiérarchie représentant un type de transactions (par exemple : les *transactions événementielles* pour une consultation ou un résultat d'analyse biologique, ... ou les *transactions persistantes* pour les antécédents, ...), une maladie (par exemple : le diabète), une unité médicale ou une période de temps fixée, ... ;
  Ces deux classes permettent de représenter la structure ou le découpage du DPI.

- La classe Entrée (**ENTRY**) pour représenter les informations médicales propement dites ; elle comprend 4 sous-classes : OBSERVATION pour représenter une observation, l'état clinique ou paraclinique du patient, ÉVALUATION (plannifcation des soins), INSTRUCTION et ACTION pour représenter les prescriptions au sens large et les actions menées en réponse aux instructions, and VIEW pour proposer des vues adaptées au besoin du DPI ;
- La classe **NAVIGATION** permet d'avoir une structuration en rubrique (bloc) des entrées. Ce concept reflète les informations collectées durant une prise en charge (par exemple : le motif d'hospitalisation, les antécédents, les allergies, les informatiosn sous la forme SOAP, ...) ;

8. Un archétype est une modélisation de concepts du domaine, exprimée sous forme de contraintes sur les données dont les instances se conforment à un ensemble de classes du RM [Beale, 2002] permettant la définition du modèle de domaine indépendamment de la terminologie utilisée

– D'autres concepts spécifiques (comme le 'versionning', la prise en compte du contexte, ...) existent [Beale et al., 2007].

Les travaux de [Cuggia et al., 2009] ont montré la limite du modèle "archétype" même si l'approche "archétype" fournit une description riche des informations (par rapport au modèle RIM). Ces informations sont principalement en format textuel et difficilement exploitables hormis par l'intermédiaire de langages et de requêtes spécifiques à OpenEHR [Ma et al., 2007].
Le modèle d'information d'OpenEHR est différent des autres standards car les spécifications OpenEHR ne sont pas limitées uniquement à la communication du DPI, il existe des spécifications pour la gestion, le stockage et la recherche (interrogation/requêtage) sur le DPI.
Il existe deux points de vues de divergence selon [Schloeffel et al., 2006] avec RIM et HISA :

1. un amalgame sémantique sur les différents systèmes existant dans les architectures de santé plutôt que sur les modèles (particulièrement les modèles d'information du DPI). RIM HL7 et HISA ont basé leur RM sur le paradigme *"d'acte (activité)"* ; tous les aspects des soins cliniques sont représentés comme des actes ;
2. RIM et HISA ne sont pas des modèles de données mais plutôt des modèles sémantiques pour spécifier les formats des messages plutôt que la structure interne des données. Dans `OpenEHR`, le DPI est vu comme un système (approche différente de celle d'HL7 basée sur les messages et plus évoluée que celle d'HISA basée sur les services). Malgré ces divergences, il y a plusieurs domaines qui doivent être harmonisés, particulièrement les types de données, les terminologies utilisées, les archétypes et templates de chaque standard.

Nous n'avons pas choisi d'implémenter ce RM, car ses spécifications ne sont pas adaptées à notre SI CISMeF. Nos données proviennent de la base CDP conforme à la norme HISA. Des équivalences peuvent se faire entre les templates HISA et les archétypes d'`OpenEHR` mais nécessitent un très grand travail d'ETL entre notre SIC vers le système OpenEHR. Par ailleurs, `OpenEHR` dispose de ses propres spécifications pour la définition et l'interrogation des données (cf. section 4.4).

4.3 Le modèle EI@DM

4.3.1 Notre approche

L'objectif de notre modélisation est de pouvoir utiliser un moteur de recherche multi-terminologique comme CISMeF [Darmoni et al., 2001b] et bientôt inter-lingue [Grosjean et al., 2011b] et des outils du Web Sémantique, particulièrement le SPARQL pour arriver à des temps de réponses minimums (de l'ordre de la seconde pour une requête sur un patient unique, et de la minute pour une requête épidémiologique sur plusieurs patients). Compte tenu de l'hétérogénéité et de la grande quantité de données qui composent un DPI, il est loin d'être facile de prendre en compte et d'aligner tous les concepts importants du dossier patient dans un modèle. Pour cela, nous avons défini 4 principes pour nous guider dans la conception de notre modèle :

- **principe 1 (P1) : Un modèle d'intégration de niveau 'type'**
  La conception doit donner la possibilité d'intégrer l'ensemble des données du DPI. Ce modèle doit inclure les données que nous avons jugées pertinentes pour notre RI : patient, prises en charge, actes médicaux, examens biologiques, courriers et compte-rendu médicaux. Sa conception doit permettre d'intégrer tout type de données sans se limiter aux données du DPI du CHU de Rouen ;
- **principe 2 (P2) : Une évolution des modèles existants**
  L'objectif n'est pas de créer un modèle *ex nihilo*, la conception doit tirer parti de manière appropriée, des travaux antérieurs sur la modélisation des données ;
- **principe 3 (P3) : Un modèle générique, flexible et sémantiquement riche**
  Le modèle doit permettre la standardisation des sources de données sur un vocabulaire commun en les mettant en relation (alignement) avec des terminologies de références (TR). Il doit permettre de suivre l'évolution des connaissances, le modèle de données ne changera pas mais uniquement le modèle de connaissances. Par conséquent, une donnée pourra être décrite et alignée vers plusieurs TR pour représenter la sémantique qu'elle véhicule ;
- **principe 4 (P4) : Un modèle centré sur la RI**
  Le modèle doit être adapté à la RI, aider l'utilisateur dans sa recherche et avoir la capacité à fournir les données dans un format adapté à la consultation par l'utilisateur.

La conception a été faite en respectant ces principes et selon les 3 approches de modélisation suivantes :

- l'approche **Entity-Attribute-Value (EAV)** pour stocker toutes les valeurs de nos métadonnées (attributs) ;

L'approche EAV [Nadkarn et al., 1999] apporte une grande flexibilité vis-à-vis du modèle de données. En effet, si le modèle vient à être étendu, il n'est pas nécessaire de modifier le schéma. Cette modélisation physique verticale permet de diminuer le temps de traitement au niveau de la base de données lors d'opérations tels que l'ajout de nouvelles métadonnées (par exemple, ajouter une nouveau type de prise en charge) au dictionnaire de données commun (ou spécifique). La représentation en ligne plutôt qu'en colonne, permet des insertions/modifications de données sans modification physique de la base de données. EAV ne pose pas de limites sur le nombre ou type d'entités, des attributs ou des relations pouvant être gérés par le modèle de données. EAV utilise largement les métadonnées, plutôt que de s'appuyer sur des structures de tables statiques et des clés étrangères pour décrire les caractéristiques et les relations entre les données d'un modèle ;

– l'approche **Entity-Relational (ER)** en tant que modèle sémantique de données ;

Le modèle ER comprend des tables avec des noms usuels pour représenter des éléments du monde réel (par exemple, les patients, les diagnostics, etc.) et des colonnes pour représenter les caractéristiques de ces éléments (le sexe du patient, le code diagnostic CIM-10, le type de prise en charge, etc.). C'est une modélisation *simple*, *intuitive*. Cette modélisation est l'approche dominante dans la plupart des SI bien qu'elle ne différencie pas clairement les informations des connaissances.

– l'approche **Data-Metadata-Semantics (DMS)** pour un modèle à plusieurs niveaux ;

L'intégration de données de complexité variée et l'évolution du domaine de connaissance nous ont conduit à adopter une approche de modélisation multi-niveau en s'appuyant fortement sur les métadonnées. Une première couche "sémantique (semantics)" pour faciliter la cohérence sémantique des données intégrées et pouvoir supporter une ou plusieurs TR. Une deuxième couche de "métadonnées (metadata)" pour représenter la structure sous-jacente des données et permettre des requêtes cohérentes au sein de populations de patients. Enfin, la dernière couche "information (data)" représentant les données à modéliser, selon un niveau de conceptualisation (niveau de granularité et d'abstraction des données) choisi.

Nous décrirons dans cette section la mise en œuvre de cette approche.

4.3.2 Conception du modèle

Pour effectuer la modélisation, nous avons analysé les données structurées et codées contenues dans le sous-modèle CDP ainsi que les données textes contenues dans les CR médicaux. Ces dernières pourront être prises en compte sous forme de métadonnées après indexation en s'appuyant sur l'univers multi-terminologique de CISMeF sur lequel se fonde notre RI [Grosjean et al., 2011a]. Nous avons aussi étudié la manière de modéliser ces métadonnées sans perdre la structure logique et sémantique du contenu (par exemple, un diagnostique X d'aujourd'hui, correspondant au motif d'hospitalisation d'un CR de séjour, devient un antécédent X d'un autre CR de séjour ultérieur). La prise en compte de ces contraintes et la mise en œuvre de cette approche de modélisation, nous ont amené à définir un modèle [Dirieh Dibad et al., 2011b] dont nous détaillerons les caractéristiques dans les paragraphes suivants.

4.3.2.1 Les données

Nous considérons le DPI comme un document contenant un ensemble d'éléments informationnels (EI) liés par des relations conceptuelles et temporelles. Le modèle d'information que nous proposons s'abstrait de la vision générale dans laquelle on perçoit un dossier médical comme un ensemble d'événements médicaux [Huff et al., 1995] ou un document temporellement riche rempli, d'affirmations au sujet de la chronologie des événements médicaux [Hripcsak et al., 2005]. Nous nous sommes fondés sur une approche semblable à celle que l'équipe CISMeF utilise pour décrire une ressource Web [Dirieh Dibad et al., 2009].
Nous considérons le dossier patient comme étant un ensemble d'EIs où chacun est décrit par des métadonnées ("attributs") et lié à un ou plusieurs vocabulaires médicaux ("descripteurs") (voir Figure 4.5). Outre la prise en compte de ces métadonnées (attributs et descripteurs), notre modèle cible devra être constitué d'un nombre de tables limité afin de minimiser les jointures implicites (patient/séjours, séjour/examens biologiques, etc.) et de permettre un traitement de requêtes simples, logiques et cohérentes. Le modèle doit pouvoir traiter au moins la sémantique qui est déjà présente dans la plupart des applications de gestion des dossiers médicaux (comme par exemple : identifier un patient, agréger des informations, etc.).
Un EI peut être une information simple concernant le patient X, un séjour d'hospitalisation pour une fracture de jambe de la date D1 à la date D2, une radiographie (acte technique) de la cheville, une prise en charge en cardiologie, un résultat d'analyse de l'hémoglobine, le compte-rendu d'un séjour (CR) ou être plus complexe : par exemple, le motif d'hospitalisation contenu dans ce CR. Ces EIs sont décrits par des métadonnées.

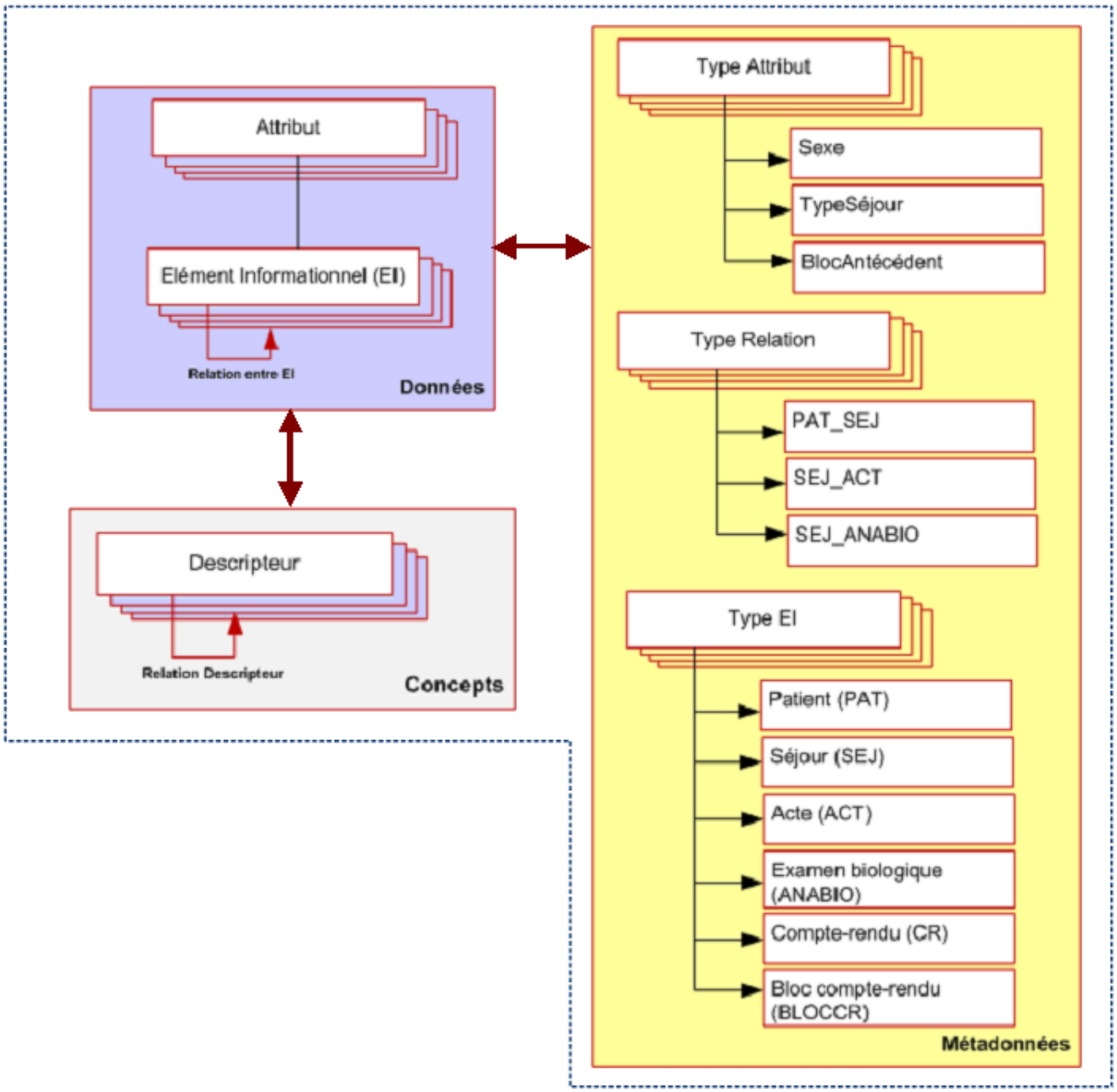

FIGURE 4.5 – Modèle à plusieurs niveaux

4.3.2.2 Les métadonnées

Les métadonnées sont utilisées pour construire des solutions d'intégration et/ou d'interopérabilité. Elles sont très importantes dans le cas de la RI. Dans notre modèle, elles définissent et décrivent précisément quelles sont les données à stocker, leur type

et comment on accède à ces données ou comment elles pourraient servir de support au traitement(voir Figure 4.6). Elles servent, aussi, à définir les relations entre ces données afin de faire des regroupements en fonction de la façon dont elles sont collectées dans un système (en l'occurrence dans le DPI).
Nous avons défini des métadonnées pour structurer les données à stocker dans notre modèle et rendre ce dernier générique et flexible. Les types d'EI sont des métadonnées pour englober l'ensemble des concepts cliniques (patient, prise en charge, CR, concept, etc.) selon le niveau d'abstraction et de granularité choisi. Ces métadonnées nous permettent de jouer sur le niveau de granularité.
Par exemple, pour les CRs, nous avons défini une métadonnée 'BLOC Compte rendu' pour décrire le contenu d'un bloc (par exemple : rechercher "par motif de recours" dans les CR, implique de stocker l'information sur le motif explicitement) ce qui est plus précis que de chercher dans tout ce CR. Nous pourrions aller plus loin en définissant des métadonnées comme celles définies pour le standard HL7 CDA et ainsi rendre notre modèle transposable vers des modèles conformes à ce standard. .
Les types d'attributs correspondent aux métadonnées caractérisant les EIs. Par exemple : le sexe pour le patient, le statut d'un séjour, les paramètres d'une analyse biologique.... D'autres métadonnées permettent de définir les relations entre les différents EIs, permettant ainsi de représenter la structuration d'origine du contenu du dossier patient, en accord avec le sous-modèle CDP (Patient-Séjour, Séjour-Acte, Séjour-Examen Biologique, ...). Chaque métadonnée possède une expression compréhensible pour l'homme et peut être traité par la machine (par exemple : `T_REL_PAT_SEJ` pour indiquer la relation Patient/Séjour, `T_INFO_PATIENT` pour regrouper les EIs par patient, `T_ATTR_TYPESEJ` pour indiquer le statut du séjour, ...), afin de présenter à l'utilisateur les types de données et l'aider et/ou faciliter le processus de formulation des requêtes.

4.3.2.3 Les concepts

Un des objectifs de ce modèle est de représenter les données cliniques par un vocabulaire standard afin qu'elles puissent être interprétées par des outils sémantiques. Les TR sont importantes pour établir un langage commun entre les professionnels mais aussi pour servir à la standardisation des données. En intégrant l'univers multi-terminologique dans notre modèle, l'alignement des données du DPI vers les TR est une étape importante de la standardisation des données.
Nous n'avons pas abordé dans cette thèse l'intérêt d'utiliser les terminologies d'interface (TI) (pour qualifier un examen biologique ou une radiologie, les professionnels n'utiliseront ni LOINC ni CCAM mais un terme adapté) et dans un second temps, ces TI pourront être utilisées pour faciliter la RI dans le DPI.

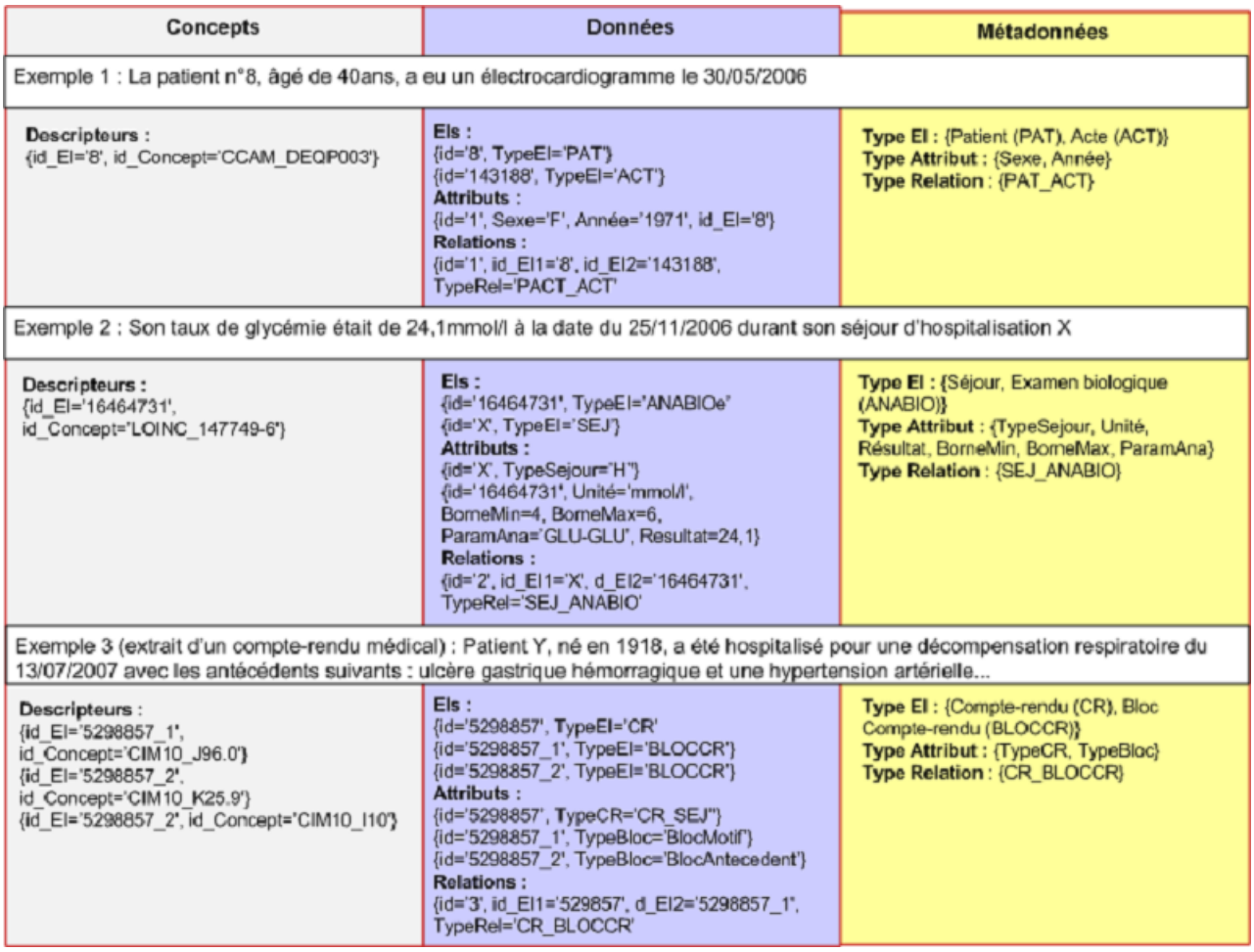

FIGURE 4.6 – Exemples de métadonnées

Les métadonnées décrivent la structure des données mais ne précisent pas comment les informations stockées peuvent être interprétées ou, comment la signification d'un sous ensemble de données peut être extrait pour permettre l'inférence de nouvelles connaissances, potentiellement contenues dans les données. Dans notre cas, cette sémantique est issue du codage PMSI pour les données codées, ou issue de l'indexation automatique des CRs à l'aide d'outils d'indexation et d'extraction sémantique de concepts (F-MTI [Pereira et al., 2009], outil de Peter Helkin en partie francisée [Sakji et al., 2011]). Pour ce qui est de l'indexation automatique des CRs, un découpage en bloc est réalisé préalablement afin de prendre en compte le contexte d'apparition des concepts (par exemple : le même concept peut apparaître dans un bloc MOTIF d'un CR puis dans un bloc ANTECEDENT dans un CR postérieur). Par ailleurs, cette sémantique est enrichie par l'univers multi-terminologique grâce aux travaux de Merabti [2010]; Merabti et al. [2011]

sur les alignements des terminologies. Ce qui offre un potentiel de raisonnement et d'inférence sur les concepts médicaux contenus dans le dossier patient ; par conséquent ces interprétations et ces inférences, réalisées à l'aide de la couche sémantique permettra aussi au DPI d'être intégré ou aligné avec des sources de données externes ou à des bases de connaissances permettant ainsi la réutilisation des connaissances comme les infobutton de [Cimino et al., 1992; Price et al., 2002; Cimino and Jianhua, 2003].
Le notion de concept correspond dans les autres modèles : au niveau "Categories" dans le modèle PEN&PAD [Rector et al., 1993], à la dimension "Concept " dans I2B2 [Murphy et al., 2006], aux sources de connaissances ontologiques "EKS concept" dans *Patient Chronicle Model* [Rogers et al., 2006], ...

4.3.3 Description du modèle

Le modèle sémantique, *(voir Figure 4.7)* que nous utilisons, est un modèle relationnel, principalement centré sur l'entité "**Element_Informationnel** (EI)". Cette entité prend en compte les informations pertinentes précédemment définies dans le paragraphe 4.1.4. L'entité "**RelationDescripteur_EI**" permet de gérer l'indexation d'un EI par un ou plusieurs concepts terminologiques (codes diagnostiques et d'actes, codes d'examens biologiques, ...). Ces deux entités contiennent l'ensemble des données du sous-modèle CDP, alors que dans ce dernier elles sont contenues dans l'ensemble des tables.
Les métadonnées spécifiques à chaque EI sont décrites par l'entité "**Attribut_EI**" qui associe à chacun des EIs, une désignation de l'attribut et une valeur alphanumérique. L'entité "**Relation_EI**" gère les relations conceptuelles et temporelles entre deux EIs et permet de garder la structure logique (liaison du séjour S au patient A par exemple) du DPI.
Les entités "**TypeInformation_EI, TypeAttribut_EI, TypeAttribut_EI, TypeRelation_EI**" constituent le squelette (ou les métadonnées) de notre modèle et permet de contrôler la cohérence et la qualité des données. En effet, la prise en compte d'un nouveau type de données nécessite la définition préalable dans ces entités de toutes ses caractéristiques et ses relations avec les données existantes. En Annexe B.1.1, vous trouverez les métadonnées de notre modèle : les types d'EI, les différents types d'attributs, les différents types de relations, ...

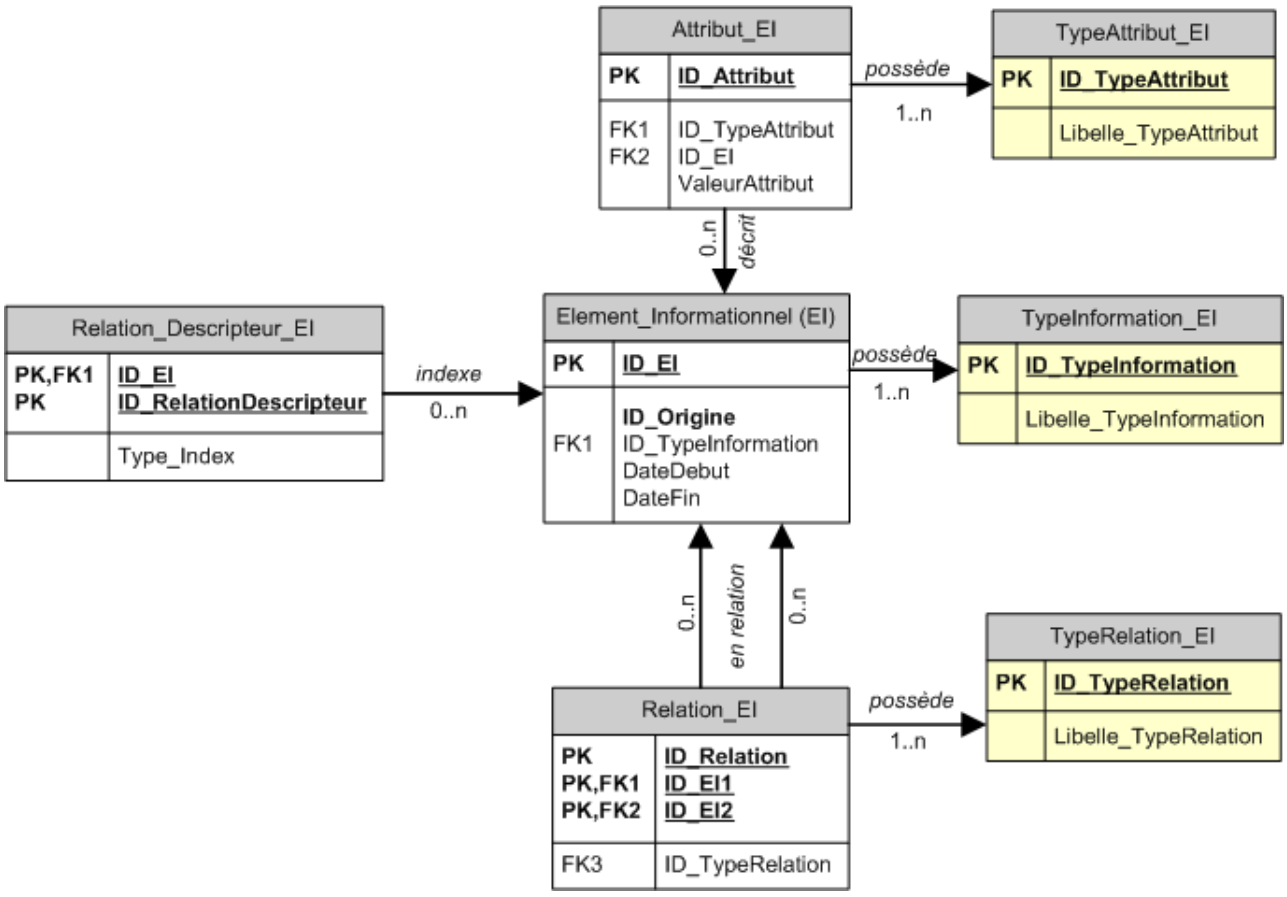

FIGURE 4.7 – Modèle RIDoPI

4.3.3.1 La portion Entité-Relation de notre modèle

Notre choix pour ce modèle a été *l'hybridation* : retenir le meilleur de chaque modèle pour aboutir à un modèle le plus concis possible, pour avoir la meilleure scalabilité possible. Nous avons combiné les 3 approches, décrites dans le paragraphe 4.3.1. Exceptée l'entité **Attribut_EI**, l'ensemble des entités constituent le point de vue logique du modèle ER, à savoir les entités **Element_Information**, **Relation_Descripteur_EI**, **TypeAttribut_EI**, **TypeInformation_EI**, **TypeRelation_EI**.

4.3.3.2 La portion Entité-Attribut-Valeur de notre modèle

Une partie des données de notre modèle a été conceptualisée dans l'approche EAV. Prenons l'exemple décrit dans le tableau 4.1 :
Cette partie de notre modèle correspond à une table de base de données unique contenant

ID_Patient	Sexe	Année_Naissance	Commune
Patient_1	F	1918	76000

TABLE 4.1 – Extrait de la table Patient

toutes les valeurs, de tous les attributs, de toutes les entités caractérisant nos données DPI. La table correspond à l'entité **Attribut_EI** de notre schéma conceptuel. Cette approche facilite le stockage et la gestion des valeurs de nos métadonnées (ajout de nouveaux attributs). Donc pour ce patient, la table EAV ressemblerait à cela (Tableau 4.2) :

ID_Attribut	EI(**Entity**)	ID_TypeAttribut(**Attribut**)	ValeurAttribut(**Value**)
1	Patient_1	Sexe	F
2	Patient_1	Année_Naissance	1918
3	Patient_1	Commune	76000

TABLE 4.2 – Extrait de la table Attribut_EI

4.3.3.3 Univers multi-terminologique et Requêtes

Les TR médicales ne se contentent pas d'activer le mécanisme de transformation des données brutes en données standardisées, elles jouent un rôle important dans la recherche et l'interrogation des données dans la base, la navigation, l'exploration du DPI à travers les hiérarchies des concepts et l'interprétation des résultats. La capacité à générer de façon efficace et efficiente des informations exploitables à partir d'une base de données

dépend de la manière dont cette terminologie est intégrée dans le modèle.
La particularité du DPI, nous avons deux types d'indexation : une indexation manuelle (MANU) issue du codage PMSI/T2A différente de l'indexation automatique (AUTO) à l'aide des outils sémantiques et/ou TAL. Nous pensons qu'il est important pour l'utilisateur de différencier si l'indexation d'un concept médical a été réalisée manuellement par un médecin ou automatiquement (par un outil). Ainsi, nous pourrons exécuter ce type de requête au sein du DPI : *"infarctus du myocarde[CIM 10][manuel] ET bilirubine > 20 ET pour la même hospitalisation.* Cette requête se traduit en langage naturel par : *rechercher tous les patients qui ont un infarctus du myocarde codé manuellement en CIM 10 et dont un dosage de bilirubine est supérieur à 20 pendant la même hospitalisation.* Ce tableau 4.3 illustre l'intérêt de stocker le type d'indexation dans notre modèle.

EI	ID_RelationDescripteur	TypeIndexation
1	CIM10_J18.0	MANU
2	CIM10_Y44.2	MANU
3	CIM10_J96.0	AUTO
4	CCAM_ZBQK002	MANU

TABLE 4.3 – Extrait de la table Patient

4.3.4 Autres avantages potentiels du modèle : Challenges du modèle

4.3.4.1 Visualisation des données

De nombreux travaux ont été réalisés sur les interfaces graphiques [9]. Notre modèle doit pouvoir être facile d'utilisation tant pour la RI au sein du DPI que pour l'exploration des données du DPI à travers des interfaces dédiées. Dans ce travail, cette problématique n'a pas été étudiée et nous ne pouvons estimer la plus-value de ce modèle pour faciliter la création de documents de synthèse du DPI comme ceux décrits dans les travaux de [Rogers et al., 2006] ou la génération de vues adaptées aux besoins des professionnels. Les méthodes utilisées pour la création de vues ou de synthèses sont principalement basées sur des requêtes, ce modèle conçu pour optimiser l'utilisation des requêtes devrait faciliter la création de vues et de requêtes, ceci sera étudié dans le cadre du projet **RAVEL**.

4.3.4.2 Interopérabilité du modèle

Comme nous l'avons décrit au chapitre 2, la base de données implémentée dans le SIC du CHU de Rouen est conforme à la norme HISA. Lors des étapes de conception, nous avons pris en compte cette norme pour permettre au modèle d'être transposable (ou être intégrer) dans d'autres systèmes conformes à cette norme. Cependant, la transposition et/ou l'intégration du modèle ne se limite pas à cette norme. Notre objectif est de permettre au modèle d'être intégré dans des (ou être support aux) outils de RI ou des outils d'aide à la décision implémentés dans d'autres systèmes. Pour cela, **`des parseurs conformes à la norme HL7`** ont été développés pour intégrer des sources de données DPI hétérogènes dans le modèle EI@DM.

9. Voir la sous section 2.2.3 du chapitre 2

4.4 Analyse comparative

Nous avons évalué dans la section 4.2.2 quelques modèles normatifs et non normatifs. Nous avons détaillé notre approche de modélisation ainsi que le modèle EI@DM. Nous allons comparer ce modèle aux autres modèles, selon des critères d'analyse résumés dans le tableau 4.4.
Dans le tableau 4.5, nous récapitulons les différentes approches de modélisation des différents modèles en spécifiant : l'élément de base de la modélisation, la méthode pour modéliser la temporalité des informations, le type de modèle utilisé, le formalisme pour représenter les données dans ce modèle, le type de base de données utilisée pour stocker les données et enfin, les méthodes pour stocker la sémantique des données dans cette base.
Dans le tableau 4.6, nous détaillons le contexte d'utilisation de chaque modèle, sa capacité à intégrer les données ainsi que des bases de connaissances, le formalisme utilisé pour rendre le modèle interopérable et enfin les outils utilisés pour interroger le modèle.

Critères	Description	Exemples
RI	Adaptation du modèle à la RI	RI texte, RI structurée, RI questions/réponses...
Intégration	Si le modèle supporte une ou plusieurs terminologies/ontologies médicales	SNOMED CT, LOINC, CIM10, CCAM, ...
	Les méthodes utilisées pour intégrer les données hétérogènes	processus d'intégration ETL, alignement de modèles, utilisation d'ontologies, utilisation d'outils TAL, modèles, ...
Interopérabilité	Les méthodes utilisées pour l'interopérabilité des données et la transposabilité du modèle vers d'autres systèmes (ou modèles)	CDA HL7, XML...
Modélisation	Le niveau de conceptualisation (granularité, abstraction, etc.)	Une observation, un événement, une valeur ou un état clinique, ...
	Les types de modèles utilisés	Un modèle relationnel, multidimensionnel, ...
	Les méthodes utilisées pour représenter la temporalité, le contexte...	Date, approche SNAP/SPAN, ...
	Les méthodes pour représenter la dimension sémantique des données	Archétypes, Dimensions, RDF schéma/OWL...
	Les outils pour stocker les données et leur dimension sémantique	BDR, BDO, XML, datawarehouse...
Interrogation	Les langages de requêtage utilisés dans ces modèles	Grid, SQL, OLAP, SPARQL, Xquery, Object-Oriented Query Language (OOQL) - *" langage d'interrogation pour les BDO"*, Archetype Query Language(AQL) - *"langage de requête déclarative pour les archétypes"*, EHR Query Language (EQL) - *""*, ...

TABLE 4.4 – Critères

	Modélisation					
	Element de base	Temporalité	Modèle		Stockage	
			Type de modèle	Formalisme	Données	Sémantique
Modèle PEN&PAD [Rector et al., 1993]	Observation	*Occurences-Level*	Modèle Objet	SMK	BD Relationnelle	-
Modèle Patient Chronicle [Rogers et al., 2006]	Evènement	Approche SNAP/SPAN	Modèle Objet	-	BD RDF	RDF Schéma/OWL
Modèle RDF [Lindemann et al., 2009]	-	Type Date	Graphe orienté	RDF	BS	RDF Schéma/OWL
HISA & RIM [Scherrer and Spahni, 1999; HL7, 2005]	Acte	Objet Date	Modèle Objet	RIM(HL7) et HISA(-)	-	Templates
OpenEHR [Beale et al., 2007]	Entry	Archétype Date	Modèle Objet	Archetype Description Language (ADL)	-	Archétypes
Modèle I2B2 [Murphy et al., 2006]	Fait	Type Date	Modèle multi-dimensionnel	-	DW	Dimensions
Modèle EIDM [Dirieh Dibad et al., 2011b]	EI	Type Date	ER+EAV	-	BD Relationnelle	Métadonnées

TABLE 4.5 – Les approches de modélisation

	Contexte		Intégration		Interopérabilité	Interrogation
	RI	Aggrégation	Terminologie(s)	Données		
Modèle PEN&PAD [Rector et al., 1993]	Non	Non [10]	Oui (Ontologies)	-	-	-
Modèle Patient Chronicle [Rogers et al., 2006]	Oui	Non	Oui	Chroniclisation	-	Interface propriétaire (SQL/SPARQL)
Modèle RDF [Lindemann et al., 2009]	Oui	Non	Oui	-	XML	SPARQL
HISA Reference Model [Scherrer and Spahni, 1999]	Non	Non	-	-	-	-
RIM Reference Model [HL7, 2005]	Non	Non	-	CDA HL7	CDA HL7	-
Modèle I2B2 [Murphy et al., 2006]	Oui	Oui	Oui	ETL	-	Interface propriétaire
OpenEHR Reference Model [Beale et al., 2007]	Oui	Non [11]	Oui	Spécifications OpenEHR	Archétypes	AQL, EQL
Modèle EIDM [Dirieh Dibad et al., 2011b]	Oui	Oui	Oui	API Java	Parseurs conformes HL7	SQL/SPARQL

TABLE 4.6 – Les approches complémentaires à la modélisation

4.5 Discussion sur le positionnement de notre modèle par rapport à l'état de l'art

Après avoir étudié les différents modèles, nous avons considéré que l'ensemble des modèles n'était pas conçu spécialement pour faciliter la RI.
Excepté les modèles RIM&HISA, les autres modèles ne sont pas des modèles d'intégration de données, ils utilisent des approches différentes, des spécifications propriétaires pour OpenEHR [Beale, 2002], des ontologies pour I2B2 [Mate et al., 2011] ou des heuristiques pour CLEF [Rogers et al., 2006] pour gérer les données.
Les modèles évalués étaient soit inadaptés au SI CISMeF soit ne répondaient pas à notre approche de modélisation. Ce qui a justifié la conception du modèle EI@DM et cela, dans une approche ER-EAV différente de celle de [Johnson, 1996]. Nous avons comparé dans ce chapitre ce modèle aux différents modèles étudiés au plan :

- de la représentation sémantique ;
- du lien entre le modèle d'information et le modèle de connaissance ;
- et de l'utilisabilité du modèle.

4.5.0.3 Représentation sémantique

Pour représenter la structure sous-jacente et la sémantique des données du DPI, certains utilisent des ontologies [Diallo, 2006; Patel and Cimino, 2007; Mate et al., 2011], se basent sur un formalisme proche de l'UML HL7 [2005], sur les archétypes Beale et al. [2007]; Beale [2002]. A travers ces différents formalismes, et particulièrement pour les modèles évalués, nous avons une représentation granulaire d'une donnée clinique différente d'un formalisme à l'autre. Le concept "Act" du modèle RIM (ou "Activity" du modèle HISA) est défini pour représenter tous les aspects cliniques afin de partager des extraits du DPI avec un niveau d'abstraction élevé.
Les actes ponctuels ou répétitifs du modèle CLEF Chronicle permettent de représenter une vue time-oriented améliorée du DPI [Rogers et al., 2006].
Le concept d'observation clinique (fait) au niveau d'une prise en charge dans le modèle I2B2 est utilisé à des fins d'analyses multi dimensionnelles [Murphy et al., 2006].
La notion d'"observation directe (un ensemble d'état clinique)" du modèle PEN&PAD permet de faire la distinction entre un fait et un opinion sur cet fait à visée d'aide à la décision et de saisie de données [Rector et al., 1993].
Le modèle doit pouvoir capturer l'information clinique de telle sorte qu'elle soit utilisable et accessible pour le besoin identifié et, par conséquent, la formalisation du modèle répond à ce besoin, en définissant le niveau de granularité de l'information clinique [Beale et al., 2007].

[Cuggia et al., 2009] abordent cette problématique et montrent que la représentation HL7 explicite les relations entre les différents éléments (le lien entre le score APGAR et le tonus musculaire) d'un concept clinique, à la différence d'OpenEHR qui liste un ensemble d'éléments dans un archétype ne comprenant pas de relations hiérarchiques entre les fragments d'archétypes. Dans le modèle EI@DM, les métadonnées peuvent nous permettre de définir des types d'observations complexes (OC) et d'observations atomiques (OA) liées par une relation d'inclusion (*OA inclus dans OC*). Dans le cas du score d'APGAR, nous aurons : *"un EI de type OC pour le score d'APGAR et 5 EIs de type OA correspondant aux actes réalisés pour calculer le score d'APGAR. Chaque EI sera décrit par un concept, par exemple le concept F-86810 pour le score d'APGAR et le concept F-11190 pour le calcul du tonus musculaire".*
Le même principe est utilisé dans les travaux de [Mate et al., 2011] mais avec une approche ontologique. Notre approche ER-EAV représente facilement ce concept. L'approche `Archétype` d'OpenEHR [Beale, 2002] permet de faire une modélisation objective en ne modélisant que des scénarios cliniques particuliers. Celle-ci repose sur une hiérarchie imbriquée de fragments d'archétypes afin d'avoir une représentation fidèle de ce qu'il est nécessaire d'enregistrer pour une observation clinique (par exemple, l'archétype "Blood pressure" contient les archétypes "Systolic" et "Diastolic"). Notre approche fondée sur les métadonnées peut correspondre, en partie, à l'approche "archétype" des spécifications d'OpenEHR, mais il n'y a pas de correspondance exacte entre nos métadonnées et les archétypes d'OpenEHR.

4.5.0.4 Lien entre le modèle d'information et le modèle de connaissance

De manière générale, la granularité est limitée par les terminologies, du fait de la nécessité de représenter chaque information par un concept, ceci crée une dépendance du modèle vis-à-vis de la terminologie. Prenons l'exemple de la représentation HL7 de la mesure de la température corporelle. Elle fait référence au code SNOMED CT 386725007 dans le bloc CDA correspondant *(<observation classCode="OBS" moodCode="EVN"><code code="386725007" ... codeSystemName="SNOMED CT" displayName="Body temperature"/><value xd :type="PQ" value="36.9" unit="Cel"/></observation>).*
Comme il n'y a pas un bloc CDA spécifique et générique à cette observation, le changement d'une terminologie vers une autre a un impact sur le modèle car l'information est codée en dur dans un document CDA.
Pour OpenEHR [Beale, 2002; Beale et al., 2007], la fonction de "Term binding" dans la modélisation multi niveau, permet de mapper un archétype (ou un fragment d'archétypes) vers une ou plusieurs terminologies standards.
Toutefois, la comparaison de la représentation du score APGAR en OpenEHR et RIM HL7 dans [Cuggia et al., 2009] montre la nécessité du choix de la bonne terminologie

(SNOMED CT plutôt que LOINC) pour bien représenter la donnée clinique. L'expressivité de la terminologie impacte la finesse de l'information à stocker.
Un des avantages du modèle EI@DM, est de diminuer ce risque de dépendance à une terminologie grâce à la séparation nette entre le modèle d'information et le modèle terminologique/ontologique.

4.5.0.5 Utilisabilité du modèle

La séparation nette entre le modèle d'information et le modèle terminologique favorise l'interopérabilité [Qamar and Rector, 2007]. Par ailleurs, la modélisation conditionne les méthodes et/ou les outils à mettre en oeuvre pour intégrer les données et les connaissances médicales dans le modèle ou pour utiliser ce dernier comme support dans d'autres applications. Certains modèles nécessitent des langages spécifiques pour interroger les données (le langage AQL et/ ou EQL pour interroger les archétypes d'OpenEHR [Beale et al., 2007; Ma et al., 2007], le SPARQL pour interroger les triplets RDF [Lindemann et al., 2009], ...), un langage d'interrogation simple SQL, étendu avec SPARQL pour combler les limites de ce langage améliore l'utilisabilité d'un modèle, et particulièrement le modèle EI@DM.

Synthèse

La complexité des données et des SI, de l'architecture de ces derniers et de leur organisation, l'inhomogénéité des SI en fonction du type d'exercice (hospitalier ou extra hospitalier en particulier), du pays ont fait privilégier selon les cas, la modélisation d'un aspect clinique (le contexte, la chronologie, ...), l'usage de telle technologie informatique ou la prise en compte d'une approche *orientée processus plutôt que patient, ...* Dans ce chapitre, nous avons procédé à une évaluation non exhaustive de quelques modèles ainsi qu'à une analyse comparative de ces modèles sur la base de critères en lien direct avec les étapes de modélisation et d'autres inhérents à ces étapes (RI, visualisation, échange et partage de données). Nous avons présenté notre approche de modélisation pour concevoir le modèle EI@DM adapté à la RI. Cette approche nous a conduit à concevoir un modèle de données générique, basé sur la notion d'EIs devant concilier un niveau de granularité suffisant pour avoir une couverture maximale du domaine (le DPI) et l'obtention d'un bon niveau de performance pour avoir une base de données performante et flexible. Dans le chapitre suivant, nous allons détaillé la mise en oeuvre de ce modèle dans deux scénarios d'implémentation pour évaluer ensuite, sa capacité à intégrer l'ensemble des données du DPI et son adaptation à la RI.

CHAPITRE 5

Mise en œuvre du modèle

Introduction

Dans le chapitre 4, nous avons présenté notre modèle EI@DM adapté à la RI. La modélisation et la recherche d'information sont les objectifs principaux de notre travail. Donc, il fallait développer des systèmes qui nous permettrait d'expérimenter le modèle EI@DM. Cette expérimentation a fait l'objet de deux scénarios d'implémentation : une implémentation dans une base sémantique à travers les technologies sémantiques d'Oracle et une implémentation dans le SI CISMeF à travers le prototype **RIDoPI** . Nous allons présenter, dans ce chapitre, ces systèmes en insistant sur les aspects technologiques. Notre approche globale d'implémentation est résumée dans la figure 5.1.

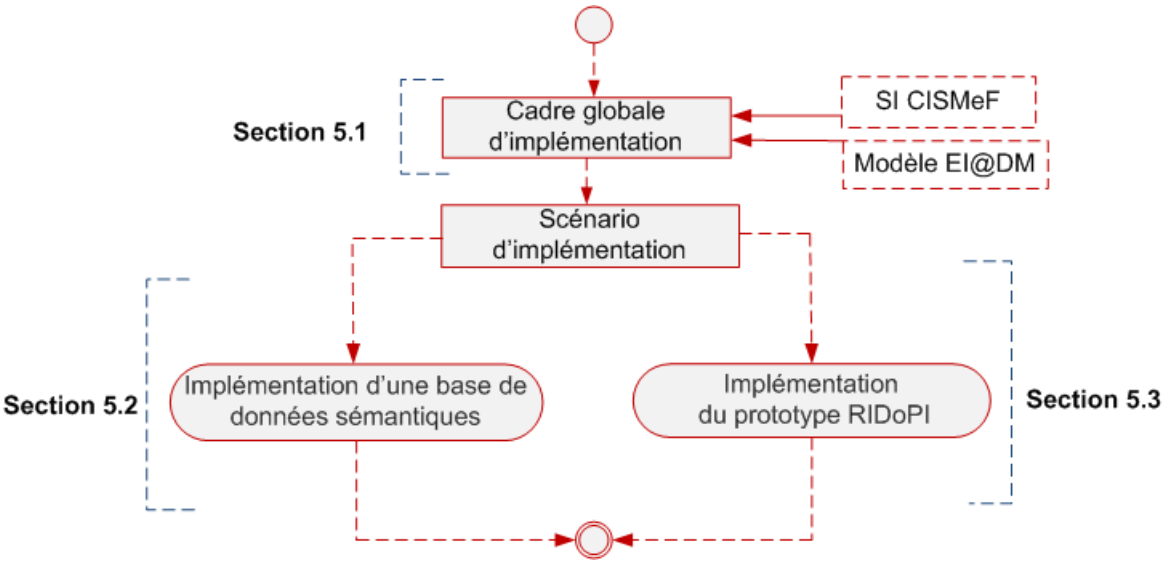

FIGURE 5.1 – Schéma synoptique de notre implémentation

5.1 Cadre d'implémentation du modèle

L'objectif de l'expérimentation du modèle EI@DM rentre dans un cadre d'adaptation des méthodes et outils existants (en particulier le moteur de recherche `Doc'CISMeF`), à l'exploitation des données DPI intégrées dans le SIC du CHU de Rouen.

5.1.1 Architecture globale

La figure 5.2 montre la place du modèle EI@DM dans le SI CISMeF. Cette architecture est composé de 4 niveaux :

1. **Query and Display Layer** correspond aux moteurs de recherches HTTP pour rechercher soit des ressources Web dans **Doc'CISMeF** (8) ou soit des données médicales de DPI dans **RIDoPI (9)**, et aux interfaces de RI pour résumer les résultats des recherches ;

   (i) **Doc'CISMeF** (Thèse de [Sakji, 2010] ;
   (ii) **RIDoPI** (cf. section 5.3).

2. **Terminology/Ontology Layer** contient un ensemble de vocabulaires standards gérés par le SI CISMeF à travers le **backoffice CISMeF (5)** et le **Portail Terminologique de Santé (PTS)(5')** ;

   (i) Le backoffice CISMeF est une base de données Oracle (version 11gR2) fondé sur un *méta-modèle* pour intégrer dans une seule STRUCTURE [1] l'ensemble des ressources Web et des ressources terminologiques (cf. sous section 5.3.1.1) ;

   (ii) Pour le PTS, un point d'accès vers un grand nombre de terminologies. Ce portail constitue une plateforme pour rassembler ces dernières dans une même structure sans se soucier ni de leur gestion ni de leur maintenance [Grosjean et al., 2011a].

3. **Data Intregation Layer** correspond aux méthodes et aux outils pour intégrer des données hétérogènes (données structurées, codées et non structurées, données issues d'autres SIC, ...) dans le modèle EI@DM parmi lesquels :

1. Voir l'Annexe B.1 pour un schéma conceptuel réduit de ce méta modèle.

(i) `Des outils TAL et/ou d'indexation sémantique pour indexer les CRH`
Cette indexation des documents médicaux s'effectue au moyen de vocabulaires structurés et standardisés (cf. niveau **Terminology/Ontology Layer**), manuellement (codage de chaque compte-rendu d'hospitalisation en CIM-10, codage de chaque compte-rendu d'acte médical en CCAM), ou au moyen d'outils d'indexation automatique développés en interne (ECMT), ou avec des partenaires industriels (F-MTI de la société Vidal) ou académiques (MCVS du Pr. Peter Elkin, Mount Sinai, NYC, USA).

(ii) `Un parseur générique PG pour répliquer les données d'une base DPI dans le modèle EI@DM :`
PG consiste en un ensemble de programmes PL/SQL de type ETL(Extract Transform Load) via une API JAVA spécifique. Ces programmes, opérationnels [2] pour notre base CDP, permettent d'extraire les données PMSI et biologiques et de les intégrer dans le modèle. Par ailleurs, ces programmes de réplication ont été adaptés pour convertir ces données en un format de transition conforme au standard HL7 [3]. Notre objectif n'est pas de détailler l'implémentation des normes CDA HL7 pour la représentation standard d'un DPI [HPRIM, 2009]. Cette adaptation a fait l'objet d'un développement de 2 autres parseurs **P1** et **P2** décrits ci-dessous :

(iii) `Un parseur P1 pour transformer un DPI en un ensemble de documents CDA HL7:`

P1 utilise des programmes PL/SQL via une API Java [4] pour extraire les données DPI et générer les parties correspondantes aux documents CDA HL7. Pour les données structurées et codées, c'est une simple réplication. Par exemple, [HPRIM, 2009] définit un type de document **SUMMARIZATION OF EPISODE NOTE** pour résumer les épisodes de soins d'un dossier patient (les séjours, les actes médicaux, les analyses biologiques, ...) [5]. Dans le cas des données non structurées, **P1** se fonde sur les outils d'indexation décrits auparavant afin d'extraire les concepts médicaux. Dans l'exemple de la figure 5.3, l'indexation du CR médical avec l'outil F-MTI [Pereira et al., 2009] a donné les concepts suviants pour le bloc "Motif" : CIM10_SC_K50.9 et SNOMEDCT_D541000 (Maladie de Crohn). Une fois ces informations extraites et conformément aux documents techniques d' [HPRIM,

2. Actuellement, nous nous fondons sur des bases de données relationnelles et **PG** sera adapté ultérieurement à d'autres SIC.
3. Documents publics HL7 - Interopsante, URL : http ://www.interopsante.org/412_p_19208/documents-publics.html
4. Les programmes PL/SQL seront adaptés aux sources de données DPI
5. Voir l'Annexe B.4 pour un extrait des épisodes de soins d'un patient au standard HL7

2009], nous générons [6] le document CDA HL7 (le bloc CDA correspondant au bloc "Motif" du CR). Nous définissons d'abord le bloc de texte (*Section-Level*) puis le bloc d'entrée (*Entry-Level*) avec le(ou les) concept(s) associés issue(s) de l'indexation.

(iiii) `Un parseur P2 pour charger des documents HL7 CDA dans notre modèle de données:`

L'exemple de la figure 5.3 montre la transposition du bloc CDA dans notre modèle. **P2** utilise ici aussi des programmes PL/SQL via une API Java pour lire les documents XML et intégrer les informations extraites dans le modèle.

6. Voir l'Annexe B.2 décrit la transformation `complète` de ce CR d'hospitalisation en un document CDA HL7

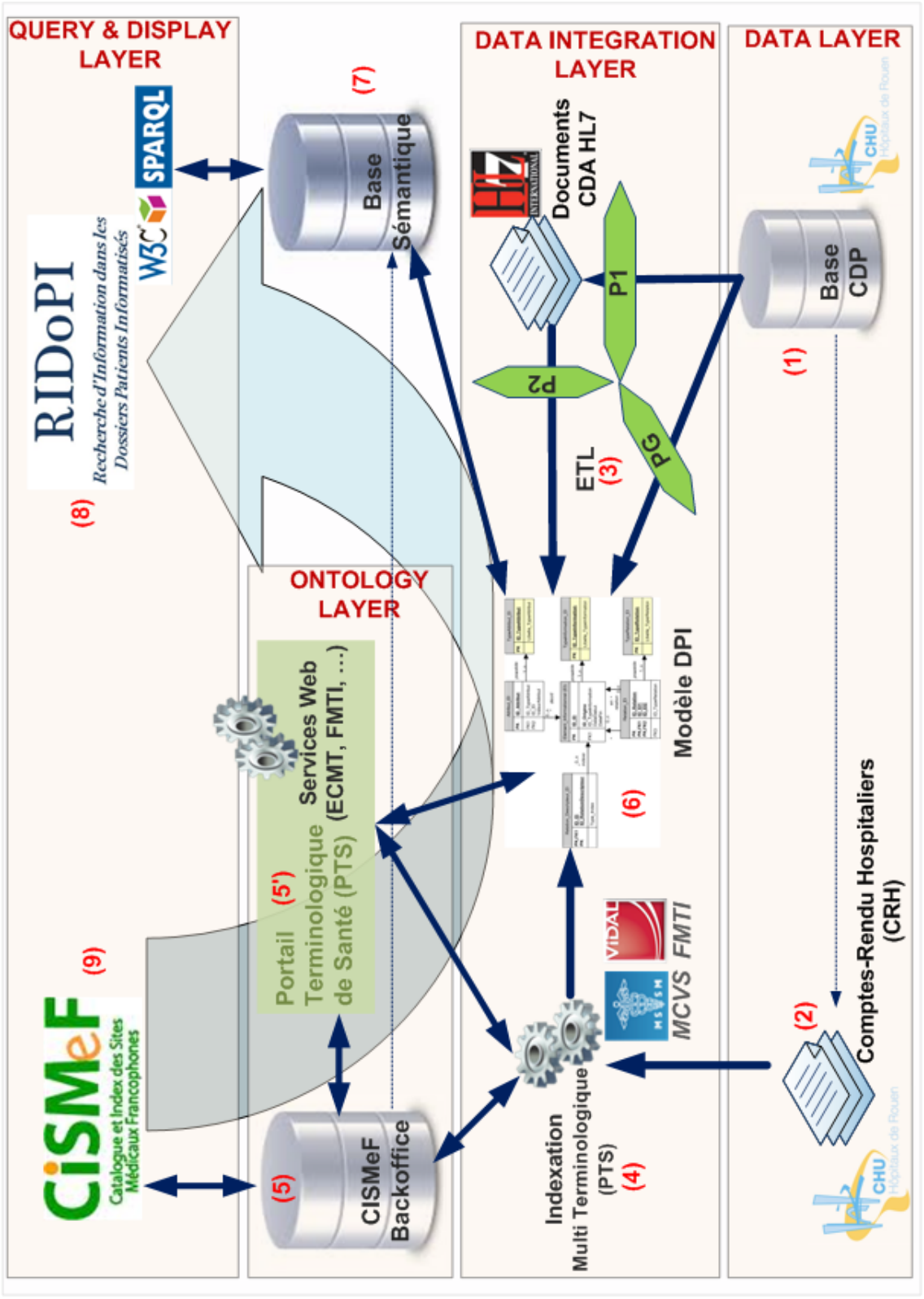

FIGURE 5.2 – Architecture globale de notre modèle dans le SI CISMeF

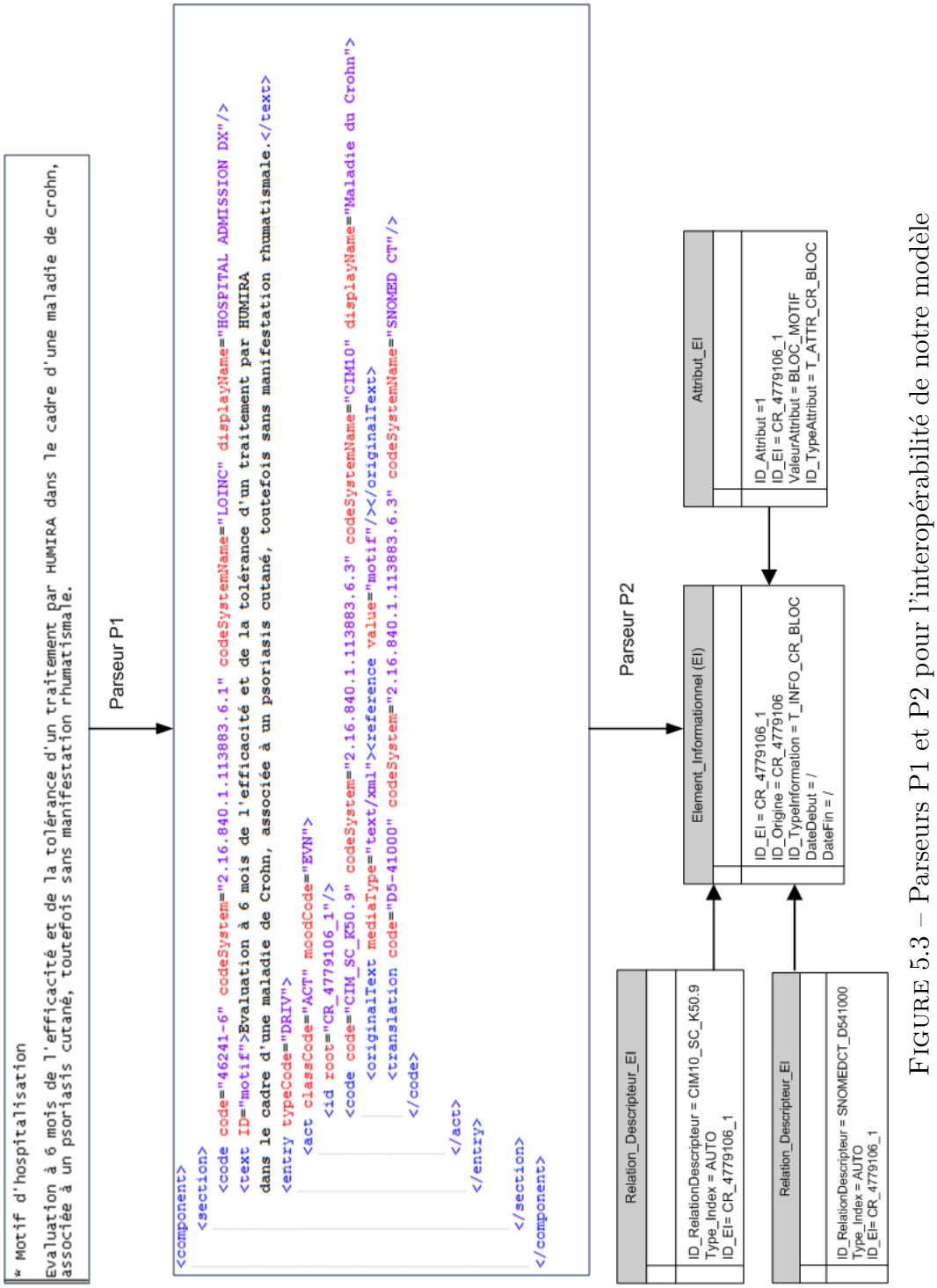

FIGURE 5.3 – Parseurs P1 et P2 pour l'interopérabilité de notre modèle

5.2 Implémentation des outils sémantiques d'Oracle

Dans cette section, nous allons décrire l'application du Web Sémantique (WS) dans le modèle EI@DM à travers les technologies proposées par la société Oracle. Ce travail, trés important dans le cadre de cette thèse de part son originalité et sa technicité, a pour but de présenter un *«preuve de concept»* de la recherche d'information au sein du DPI, en utilisant les technologies sémantiques et en particulier le langage SPARQL.

5.2.1 Web Sémantique et Oracle

Le projet `Web Sémantique` est de rendre l'information accessible et compréhensible non seulement pour les humains mais aussi pour les machines. Le Consortium du World Wide Web (W3C), chargé de développer des nouvelles technologies, propose une extension de la notion des métadonnées utilisables par les machines en définissant un cadre général de standardisation de ces métadonnées sur le Web. Dans le cadre du Web Sémantique, W3C valide un langage standard "Resource Description Framework (RDF)" pour définir ces nouvelles métadonnées.

5.2.1.1 RDF et RDFS

RDF est un formalisme basé sur un modèle sémantique de graphes étiquetés et orientés. Ce modèle représente la grande nouveauté du WS pour décrire la sémantique d'une ressource (par exemple ces caractéristiques, comme le typage de certains termes et expressions) en définissant des métadonnées formelles. Ces métadonnées sont sous la forme d'un triplet (S, P, O) composé de trois éléments :

1. **Sujet (S)** : cet élément représente la ressource à décrire, nécessairement identifiée par une URI [7] ; une ressource peut être un document, une donnée, ... ; la ressource correspond aux noeuds du graphe RDF ;
2. **Prédicat (P)** : cet élément correspond à une propriété pour caractériser et décrire **S**, nécessairement identifiée par une URI ; les propriétés correspondent aux arcs étiquetés du graphe RDF ;
3. **Object (O)** : cet élément représente une donnée de type primitive (numérique, littérale, ...) ou une autre ressource S identifée par une URI ; elles correspondent aux feuilles du graphe ou à un autre noeud.

7. Uniform Resource Identifier

Un ensemble de ces triplets forment ce qu'on appelle `"un graphe RDF"`. RDF utilise le vocabulaire RDF Schema (RDFS). Ce dernier est un langage extensible de représentation des connaissances pour fournir des éléments de base à la définition de vocabulaires destinés à structurer les ressources RDF. Nous pouvons citer par exemple le vocabulaire FOAF [8] contenant un ensemble de classes pour décrire des personnes et les relations qu'elles entretiennent entre elles (affiliations).
W3C propose plusieurs syntaxes pour décrire un graphe RDF permettant une abstraction au niveau de l'expression et une meilleur interopérabilité. Les principales syntaxes sont : le langage RDF/XML [9], la notation N-Triples [10], la notation Turtle [11] plus légère, la notation N3 [12] plus compacte. Toutefois, ce modèle RDF est limité quant à la description des contraintes sémantiques et des raisonnements ou en ce qui concerne la définition intentionnelle de concepts (impossible de définir des classes de concepts complexes avec RDFS). D'ou la nécessité d'utiliser d'autres formalismes de plus haut niveau tels que les ontologies (voir sous section 5.2.1.2).
Dans ce scénario d'implémentation, nous utilisons ce modèle RDF pour représenter les terminologies médicales (cf. paragraphe 5.2.2.2).
La figure 5.4 correspond à la représentation graphique du `concept F45.33` de la classification CIM10. Le Serveur Multi Terminologique de Santé (SMTS) [Merabti, 2010], réalisé par le partenariat entre la société MONDECA [13], CISMEF et le LERTIM, fournit la majorité des identifiants uniques (URI).
Par exemple, l'URI *<http ://www.mondeca.com/system/publishing#level>* permet de définir le niveau hiérarchique du concept dans la classification CIM10, "<http ://www.chu-rouen.fr/smts#CIM10_F45.33>" est l'URI pour identifer le concept "F45.33", "<http ://www.chu-rouen.fr/smts#CIM10Type>" est l'URI pour spécifier la nature du concept (Diagnostic, ...), ...
Ce graphe RDF [14] peut être analysé différement pour mettre en relief le couple (Prédicat, Objet) appliquée à un Sujet, en d'autres termes une métadonnée (ici notre concept F45.33) et ses caractéristiques :

- F45.33 **(S)** est un concept CIM10 de niveau **(P)** "subdivision" **(O)** ;
- F45.33 **(S)** est un concept CIM10 de type **(P)** "*Diagnostic(D)*" **(O)** ;
- F45.33 **(S)** a pour libellé en français **(P)** l'expression "dysfonctionnement neuro-végétatif somatoforme|Système respiratoire **(O)**" ;

8. Friend of a Friend, URL - http ://rdfweb.org/foaf/
9. RDF/XML Syntax Specification, URL : http ://www.w3.org/TR/REC-rdf-syntax/
10. N-Triples, URL : http ://www.w3.org/2001/sw/RDFCore/ntriples/
11. Terse RDF Triple Language, URL : http ://www.w3.org/TeamSubmission/turtle/
12. Notation 3 Logic, URL : http ://www.w3.org/DesignIssues/Notation3.html
13. Mondeca, URL : http ://www.mondeca.com
14. Voir en Annexe A.3, les descriptions (RDF/XML, N3) correspondant à ce graphe RDF

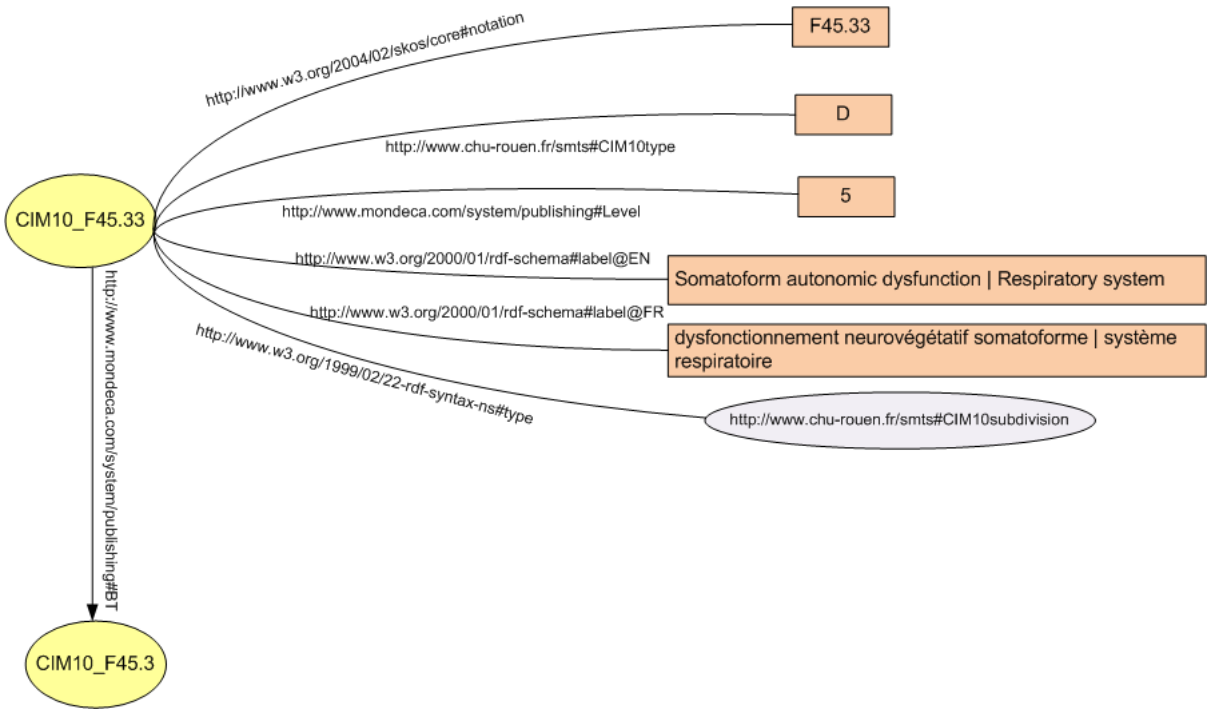

FIGURE 5.4 – Un extrait du graphe RDF du concept CIM10_F45.33

- F45.33 **(S)** a pour libellé en anglais **(P)** l'expression "Somatoform autonomic dysfunction|Respiratory system **(O)**" ;
- F45.33 **(S)** est un concept CIM10 de niveau **(P)** 5 (O) ;
- F45.33 **(S)** a pour père **(P)** le concept F45.3 **(O)** ;
- ...

5.2.1.2 Ontologies et OWL

Dérivé des projets américains DAML (DARPA Agent Markup Language) et européens OIL (Ontology Inference Layer), OWL(Web Ontology Language) est proposé par le consortium WWW (W3C) comme une *"extension"* de RDFS. OWL est basé sur une sémantique formelle définie par une syntaxe rigoureuse [15]. Il existe trois versions du langage : *OWL Lite, OWL DL, et OWL Full.*
OWL offre un vocabulaire riche pour la description d'ontologies complexes afin de faciliter l'expression de relations complexes entre différentes classes RDFS, ainsi que l'expression

15. OWL Web Ontology Language Overview, URL : http ://www.w3.org/TR/owl-features/

de contraintes plus précises sur des classes et des propriétés spécifiques.
OWL et RDFS sont tous deux des modèles RDF permettant de définir des vocabulaires. L'utilisation d'ontologies et de RDF au sein d'une application (par exemple une base de données) implique la disponibilité d'un langage d'interrogation fondé sur la structure des triplets et la sémantique des vocabulaires OWL et RDFS. C'est le rôle joué par SPARQL (Protocol and RDF Query Language).

5.2.1.3 SPARQL

SPARQL exploite l'approche sémantique des données RDF. SPARQL est doté d'un langage de requête fondé syntaxiquement sur les triplets, d'un protocole d'accès comme un service Web (SOAP) et d'un langage de présentation des résultats au format XML.
La figure 5.5 représente l'architecture de SPARQL en 3 niveaux :

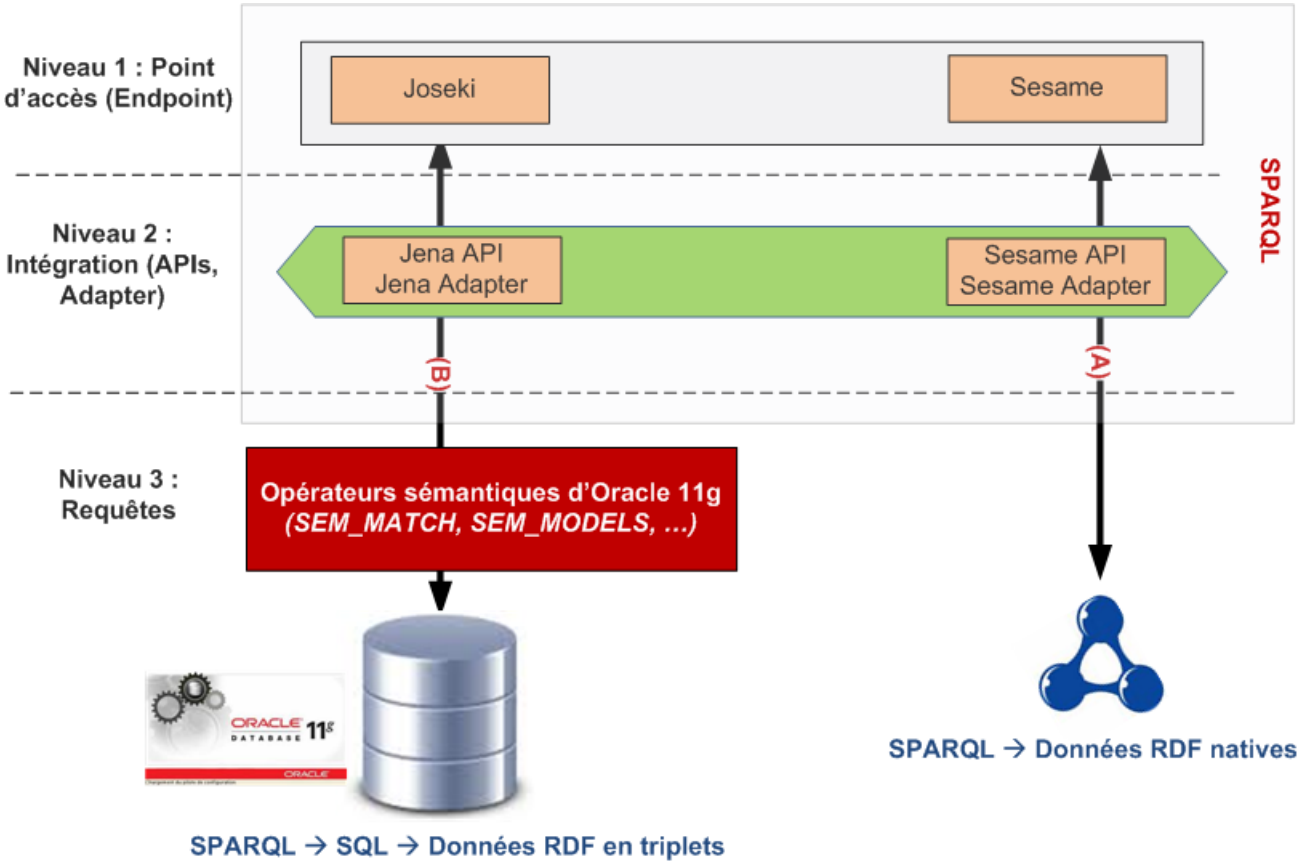

FIGURE 5.5 – Architecture SPARQL

- **au** 1^{er} **niveau** : nous avons les *"SPARQL Endpoint"*, des points d'entrées qui supportent le protocole et le langage SPARQL afin d'interroger des données RDF. Plusieurs solutions coexistent à la fois des applications clientes SPARQL et des applications HTTP SPARQL (Joseki, Sesame [16]) et ils sont dépendants du $2^{ème}$ **niveau**. Notre choix s'est tourné vers **ce moteur de recherche HTTP Joseki**. La collaboration avec Oracle nous a permis d'implémenter Joseki afin d'accéder et d'interroger les modèles RDF stockés dans une base de données relationnelles d'Oracle dans sa version 11g [17] ;
- $2^{ème}$ **niveau** : nous avons les *"Frameworks"* permettant la gestion (création, modification, suppression, insertion, ...) des graphes RDF comme celui de la figure 5.4, stockés dans un format de stockage donné (une base de données, un fichier plat, ...). Ils en existent plusieurs pour les différents langages de programmation : Jena [18] pour Java, RAP [19] pour PHP, Redland [20] pour C et RDFlib [21] pour PYTHON. `Le moteur Joseki` implémente l'API Jena et son adaptateur pour manipuler les modèles RDF.
- $3^{ème}$ **niveau** : les deux niveaux précédent constituent l'architecture *NATIVE* de SPARQL et correspondent au cas **(A)** dans la figure 5.5 . SPARQL permet d'interroger des graphes RDF `matérialsés` [22]. A l'inverse, le cas **(B)** montre que SPARQL peut interroger des graphes RDF `virtuels` [23]. Nous décrirons dans la section 5.2.1.4 les principes de mise en oeuvre de cette base de données sémantique.

A partir du graphe RDF de la figure 5.4, grâce au moteur Joseki, nous pouvons, par exemple, obtenir la description détaillée du concept F45.33 (*Requête 1*) ou bien déterminer tous les concepts CIM10 dont le métaterme associé est "infectiologie" ainsi que leurs fils (*Requête 2*). D'abord, nous définissons ces préfixes pour éviter de les réecrire dans les requêtes :

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX publishing:<http://www.mondeca.com/system/publishing#>
```

16. Sesame - un serveur open-source de stockage, d'interrogation et d'inférence des données RDF (URL : http ://www.openrdf.org/)
17. Grâcieusement prêtés par cette société dans le cadre du projet européen PSIP
18. Jena Semantic Web Framework, URL : http ://jena.sourceforge.net/index.html
19. RDF API for PHP, URL : http ://www4.wiwiss.fu-berlin.de/bizer/rdfapi/
20. Redland RDF Librairies, URL : http ://librdf.org/
21. RDFLib, URL : http ://www.rdflib.net/
22. SPARQL permet d'interroger n'importe quel composant d'un triplet qui a la forme **Sujet-Prédicat-Objet**. On n'a pas besoin de connaître a priori la structure et le contenu des données pour pouvoir les interroger.
23. des graphes RDF stockés dans une base de données relationnelle.

```
PREFIX smts: <http://www.chu-rouen.fr/smts#>
```

Requête 1 :

```
SELECT ?p ?o
WHERE {
smts:CIM10_F45.33 ?p ?o .
}
```

Requête 2 :

```
SELECT ?fils
WHERE {
?rel rdf:type publishing:BT-NT .
?rel publishing:BT ?pere .
?rel publishing:NT ?fils .
?fils rdfs:label ?label .
?pere smts:CIM10MT "infectiologie" .
FILTER (lang(?label) = "fr")
}
```

Nous pouvons réexécuter la *Requête 2* dans la *Requête 3* en utilisant une requête SQL. Cette dernière va inclure une syntaxe SPARQL en utilisant les opérateurs sématiques d'Oracle : **SEM_MATCH** pour intégrer le graphe RDF à rechercher, **SEM_MODELS** pour définir le modèle RDF sur lequel comparer le graphe RDF. La *Requête 3* interroge UNIQUEMENT des triplets RDF

Requête 3 :

```
SELECT fils
FROM TABLE (
SEM\_MATCH('
( ?rel <rdf#type> <publishing#BT-NT>).
( ?rel <publishing#BT> ?pere).
( ?rel <publishing#NT> ?fils).
( ?fils <rdfs:#label> ?label) .
( ?pere <smts#CIM10MT> "infectiologie").',
SEM\_MODELS('CIM10'),null,null,null)
)
WHERE label$RDFLANG = 'fr'
```

Nous pouvons aussi interroger des tables en même temp que les triplets RDF, ce qui permet de déterminer l'ensemble des données du DPI indexés par les concepts CIM10 résultants de la *Requête 3* (cf. section 6.1.2.1).
Les résultats de l'exécution des requêtes SPARQL (1 et 2) dans Joseki sont disponibles en Annexe A.4.

5.2.1.4 Base sémantique d'Oracle

Dans ce paragraphe, nous ne détaillerons pas l'architecture technique de la base de données Oracle dans sa version 11gR1 [24] mais nous donnerons une bref description fonctionnelle de ce qui constitue notre base de données sémantique. Oracle 11gR1 possède un moteur sémantique permettant de manipuler facilement des triplets RDF. Ce moteur sémantique intègre le point d'accès `Joseki` comme Interface-Homme-Machine. La mise en oeuvre de ce moteur nécessite un serveur d'application Java : Weblogic [25] est le meilleur candidat. La figure 5.6 résume l'architecture fonctionnelle de la base sémantique d'Oracle. 4 parties composent cette architecture :

1. la partie **"REQUÊTES"** permet d'interroger les données de la base par deux méthodes ;
2. la partie **"DONNÉES"** permet d'intégrer simultanément, dans une seule base, des données au format relationnel et des données au format de triplet RDF. Il est possible d'implémenter la représentation graphique des triplets sous une forme relationnelle [26] où la table contient trois colonnes correspondant respectivement aux triplets **S, P, O** ;
3. la partie **"INFÉRENCES"** permet d'appliquer des processus de raisonnements sur les données RDF. Ces raisonnements [27] se fondent sur des bases de connaissances et/ou sur des règles afin d'en déduire des nouvelles informations ;
4. la partie **"STOCKAGE"** correspond aux méthodes et aux outils permettant d'intégrer des triplets RDF dans la base sémantique.

24. Oracle Database Semantic Technologies Developer's Guide - 11g Release 1 (11.), URL : http ://download.oracle.com/docs/cd/B28359_01/appdev.111/b28397/sdo_rdf_newfeat.htm
25. Un autre produit d'Oracle fourni dans le cadre de notre collaboration avec la société Oracle
26. Le modèle RDF est analogue au modèle relationnel
27. Plusieurs moteurs d'inférences gratuits ou commerciaux existent, tels que Racer, Pellet, Fact, Fact++, Surnia, F-OWL et Howlet

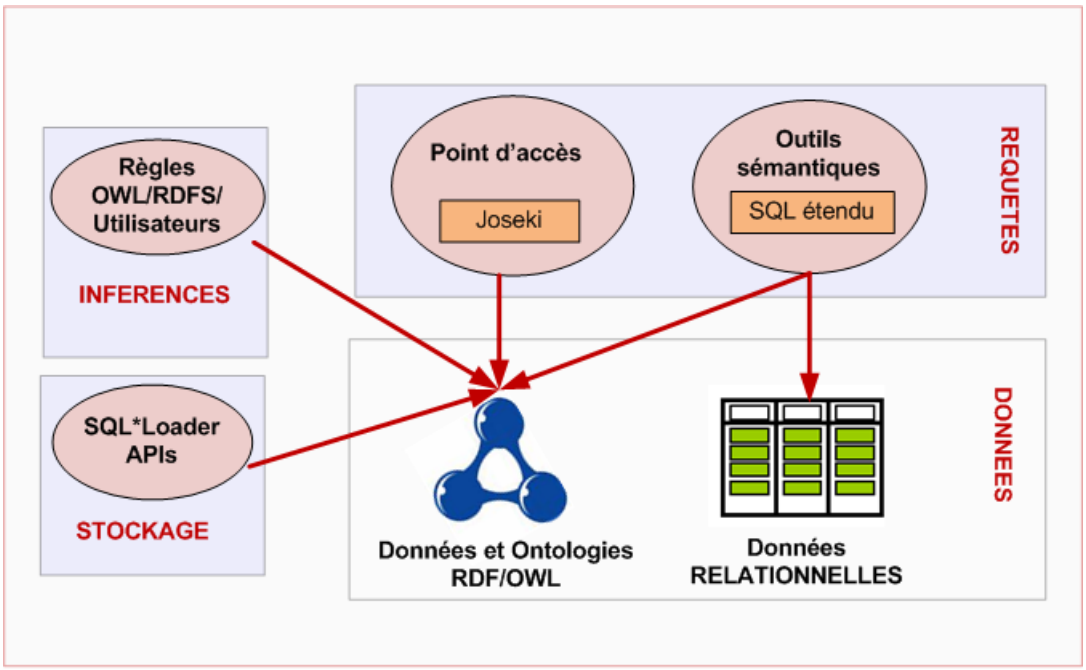

FIGURE 5.6 – Architecture fonctionnelle simplifiée de la base sémantique d'Oracle 11gR1

5.2.2 Principes de notre implémentation

Ainsi, nous avons conçu une base de données sémantique sous Oracle 11gR1 [28] en créeant dans un premier temps, le modèle de données générique pour stocker les données DPI au format relationnel, et dans un second temps, les modèles sémantiques pour stocker les données terminologiques au format RDF.
L'implémentation du modèle EI@DM dans cette base sémantique a fait l'objet d'un article [Dirieh Dibad et al., 2011b] et d'un poster [Dirieh Dibad et al., 2011a]. L'évaluation de cette implémentation sera faite dans la section 6.2.1 du chapitre 6.

5.2.2.1 Replication des données DPI

Après la construction de la base sémantique, nous avons sélectionné [29], avec l'expertise du Dr P. Massari, une vingtaine de DPI. Nous avons répliqué uniquement les données structurées (données démographiques, données de prises en charge hospitalières, codage

28. RDF Semantic Data Management Using the Oracle Spatial 11g Option. URL : http ://www.oracle.com/technology/obe/11gr1_db/datamgmt/nci_semantic_network/nci_Semantics_les01.htm
29. Cette sélection a été faite selon ces critères suivants : "*Age du patient > 40 ans et (Nombre minimum d'hospitalisations dans le CHU > 2 ou au moins un séjour en hospitalisation d'une durée > 3 semaines ou nombre de documents contenus dans le dossier du patient >= 15)*"

des diagnostics et des actes (PMSI) et résultats d'analyses biologiques) de la base CDP correspondant à ces dossiers et conformes au sous-modèle CDP réduit, directement dans notre base sémantique grâce au parseur générique. Cette étape d'extraction et de chargement s'est faite de manière semi-automatique (données pré-sélectionnées manuellement et chargement automatique grâce à ce parseur). Les données du DPI sont anonymisées en remplaçant les noms et prénoms par des séquences de caractères (Patient1, Patient2, ...). Outre ces données, nous avons aussi stocké dans notre modèle tous les identifiants nous permettant d'accéder directement à la base CDP. Des référentiels ont été définis pour certaines données (liste des modalités des séjours, liste des unités médicales, liste des types de séjours, ...). La base sémantique compte sur ces 20 DPI, un total de 2075 prises en charge, 2377 actes médicaux et plus de 20 000 analyses biologiques. Nous illustrons une première instanciation de la replication des données CDP vers notre modèle dans la figure 5.7. Dans cette figure, nous décrivons la patiente n°8 pour laquelle un électrocardiogramme (ECG) a été effecuté avant et après la destruction d'un foyer arythmogène atrial, lors d'un même séjour d'hospitalisation.

5.2.2.2 Application du modèle RDF pour deux classifications : la CIM10 et la CCAM

Bien que notre modèle puisse gérer plusieurs terminologies médicales et en fait toutes les terminologies et ontologies inclus dans notre SI (soit actuellement n = 32) pour indexer le DPI, nous nous sommes limités dans ce scénario d'implémentation aux classifications CIM-10 et CCAM. Nous avons chargé ces données terminologiques dans la base sémantique en plusieurs étapes (voir figure 5.8) :

- 1er étape : elle consiste à disposer de ces données au format OWL. Cela a été possible grâce au SI CISMeF qui permet d'exporter ces données terminologiques au format OWL ;
- 2e étape : nous avons transformé ses données OWL en format de triplets RDF et les avons chargés directement dans la base sémantique via un programme Java que nous avons développé. Ce programme utilise l'API Jena pour se connecter à la base sémantique et pour y insérer les triplets.

5.2.2.3 Processus RI

Dans ce scénario d'implémentation, le processus de RI est schématisé par la figure 5.9. Les différentes étapes de ce processus sont :

- La recherche manuelle et la validation par l'expert du domaine des termes qui seront contenus dans les requêtes SPARQL ; les expressions *"Electrocardiogramme,*

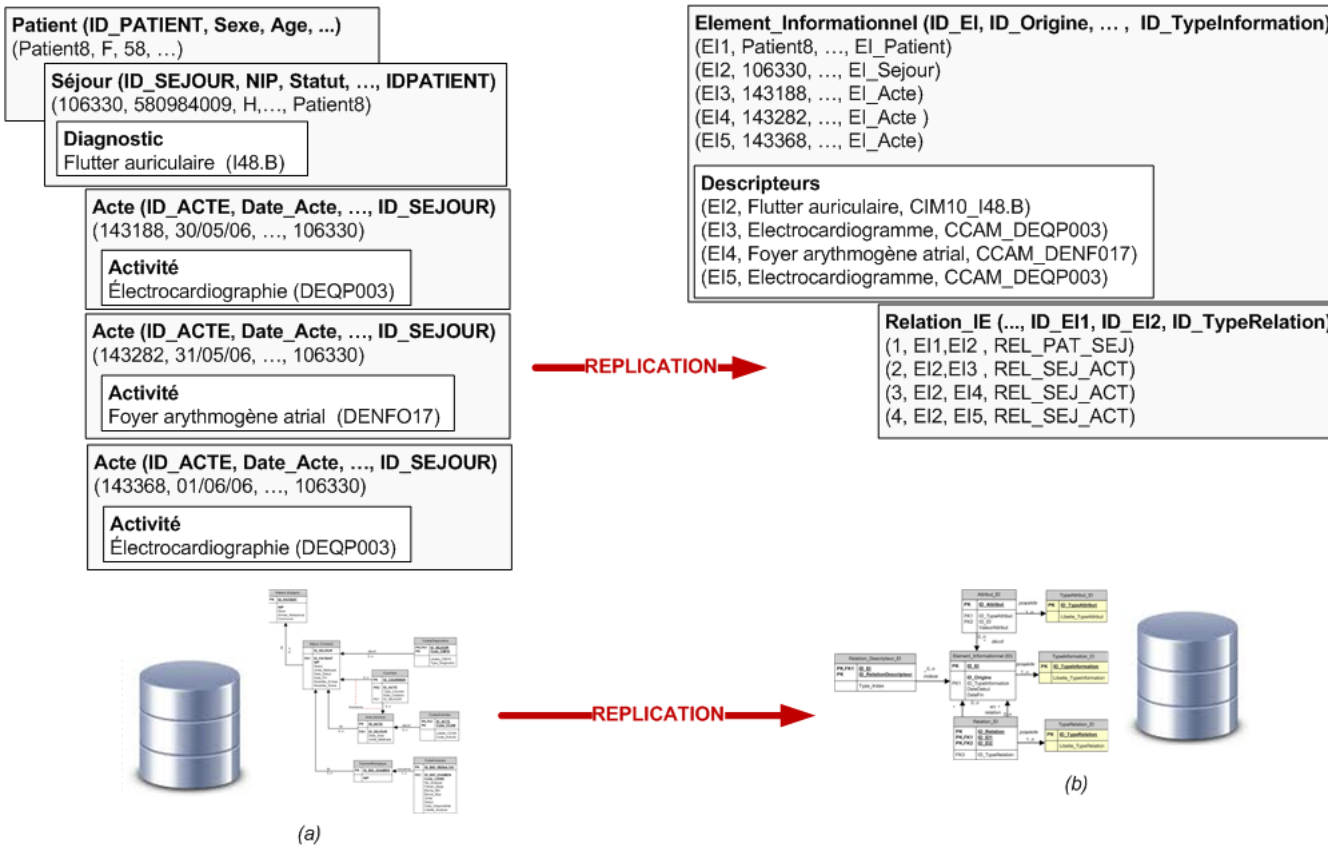

FIGURE 5.7 – Réplication des données du sous modèle CDP (a) vers le modèle EI@DM (b)

ECG" correspondent aux termes de la requête pour la question clinique *"Rechercher le dernier électrocardiogramme d'un patient X"* ;

- Le mapping automatique, des termes de RI vers les termes d'indexation (ou descripteurs) via des requêtes SPARQL sur les données RDF. Ces termes d'indexation sont représentés par un sac de mots de terminologies médicales ;
- La détermination des éléments informationnels indexés par les termes de ce sac de mots.

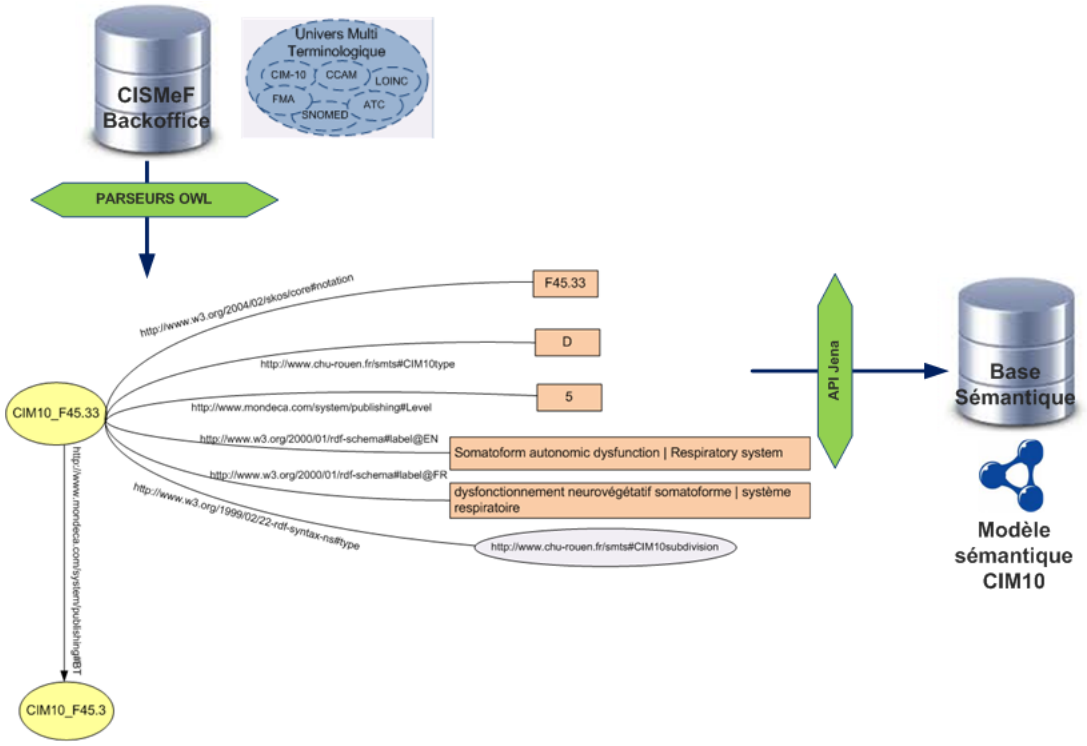

FIGURE 5.8 – Transformation des données terminologiques du SI CISMeF vers la base sémantique

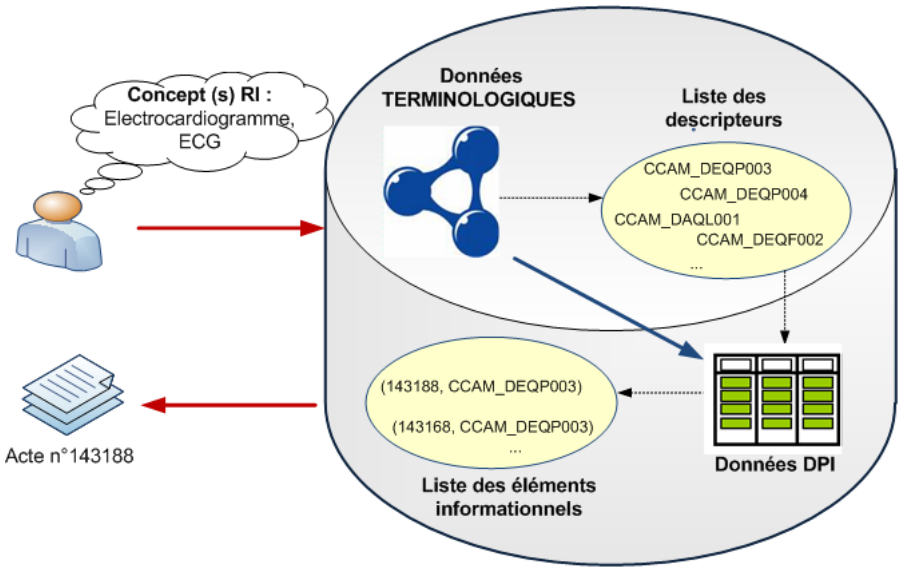

FIGURE 5.9 – Etapes de la RI

5.3 Prototype d'interfaces de RI : RIDoPI

Dans cette section, nous décrirons, sans rentrer dans les détails techniques, l'implémentation du modèle EI@DM dans le **SI CISMeF**. Cette implémentation a fait l'objet du développement d'un prototype **RIDoPI**. Cette implémentation a fait l'objet d'un article [Griffon et al., 2011]. Son évaluation sera faite dans la section 6.2.2.2 du chapitre 6.

5.3.0.4 Objectifs

RIDoPI a pour objectif de répondre au besoin d'accéder à une information médicale contenue dans un grand gisement de données. Notre prototype se comporte comme un moteur de recherche support au SIC afin de permettre l'édition d'un maximum de requêtes sur les données du DPI répondant ainsi aux besoins informationnels des professionnels qui veulent un outil d'aide à la décision dans le processus de prise en charge du patient. Les objectifs de ce prototype sont principalement :

- d'avoir une interface unique pour faire de la RI ;
- de développer un outil de RI indépendant du SIC qui gère les données du DPI ;
- de pouvoir intégrer ce prototype dans un autre SIC et faciliter la RI

5.3.1 Architecture de fonctionnement

Ce prototype utilise comme support l'architecture technique déjà disponible dans le SI CISMeF (cf. sous section 5.1.1).

5.3.1.1 Intégration du modèle EI@DM dans le modèle CISMeF

L'équipe CISMeF a conçu un méta-modèle [30] générique et flexible pouvant encapsuler n'importe quel modèle et comportant des règles (nommage, domaine, structure de données, ...) pour en assurer la cohérence et la qualité des données. Ce méta-modèle se compose en 2 parties : **"le modèle"** pour définir et enregistrer les règles et les métadonnées sur les données à gérer et **"l'intance du modèle"** pour stocker les données selon les règles et les métadonnées définies. Dans le SI CISMeF, ce méta-modèle intègre l'ensemble des ressources terminologiques et des ressources documents indexées. Le moteur de recherche DocCISMeF est fondé sur ce modèle.
Le développement de ce prototype se fonde aussi sur ce méta-modèle, donc une étape d'intégration du modèle EI@DM vers ce méta-modèle a du être réalisée.

30. Voir Annexe B.1 pour le méta modèle complet.

Les éléments spécifiques à notre modèle (règles, contraintes, type d'entité, type d'attribut, type de relation, etc.) ont été défini au sein de ce méta modèle. L'utilisation d'un méta modèle permet aux données et aux terminologies d'être indépendants et génériques, puisque le modèle terminologique et le modèle EI@DM sont des instances.
Cette intégration (voir Tableau 5.1) a été facilite par la convergence du modèle EI@DM vers ce méta modèle.
Pour cette intégration, nous avons pris comme exemple le concept "SEJOUR" [31]. L'intégration se fait en deux étapes : la $1^{ère}$ consiste à définir les éléments suivants pour ce concept : **DM_STAY** pour représenter le concept, **DM_ID_STAY** pour représenter un attribut du concept correspondant à son identifiant dans la base CDP ou **DM_IS_STAY** pour définir la relation "PATIENT-SEJOUR", et la $2^{ème}$ étape consiste à charger les données CDP dans le SI CISMeF en utilisant le parseur générique **PG**.

Modèle EI@DM	**Méta modèle CISMeF**
Element_Informationnel	**TB_OBJECT**
Attribut_EI	**TB_DATATYPE_PROPERTY**
Relation_EI	**TB_OBJECT_PROPERTY**
Relation_Descripteur_EI	**TB_INDEXING**

TABLE 5.1 – Converges des modèles

5.3.1.2 Présentation des interfaces de RIDoPI

Le prototype **RIDoPI** permet de faire de la RI au niveau d'un dossier individuel [**RI mono patient**] pour retrouver en quelques secondes *"le compte-rendu de la (ou des) échographies d'un patient déjà hospitalisé de nombreuses fois au CHU"* ; à l'échellon d'un service ou de l'hôpital [**RI multi patient**], à des fins d'analyses en épidémiologie ou en recherche clinique afin de retrouver *"tous les patients d'un service ayant déjà un infarctus du myocarde et une troponine > x ou d'envoyer automatiquement au bon investigateur le fait qu'un patient récemment hospitalisé, éventuellement dans un autre service, répond au critères d'inclusion d'une etude clinique"*.
Le prototype est en cours de développement : de nombreuses fonctionnalités ne sont pas complétement opérationnelles (comme le *post* traitement des résultats d'une réquête utilisateur [32]). La forme définitive des interfaces de RI n'est pas non plus fixée. Trois fonctionnalités de l'outil RIDoPI seront présentées dans les paragraphes suivants.

31. Voir Annexe B.1.3, l'intégration complète du concept "SEJOUR" dans le méta modèle CISMeF.
32. Nous donnons les résulats en vrac

5.3.1.2.1 L'interface RI mono patient

L'interface de RI au sein d'un dossier médical est divisée en deux parties : une première partie pré-rempli avec les données d'identification du patient et ses caractéristiques (nom, prénom, date de naissance, sexe) tandis que la seconde est constituée de trois onglets, de complexité croissante. Ces onglets permettent de faire des recherches complexes nécessaire au formalisme des fonctionnalités de recherche. Entre autre, elles incluent des opérateurs booléens (ET, OU, ...), des opérateurs temporelles (AVANT, APRES,...), des recherches sur des métadonnées (concepts, valeurs numériques, ...)(cf. Figure **??**). A travers cette interface **RI mono patient**, l'utilisateur peut effectuer une interrogation séparée ou intégrée sur les actes et prises en charges hospitalières et les données biologiques. Cette interface se compose de 3 onglets :

- Le 1er onglet (Figure 5.10) permet d'effectuer des recherches sur les termes d'indexation [33] ainsi que sur un nombre restreint de métadonnées (Unité médicale, Durée, Date d'entrée et Date de sortie). Possibilité d'associer plusieurs conditions à l'aide des opérateurs Booléens et de préciser si ces différentes conditions doivent concerner un seul séjour (*même hospitalisation*). Sans détailler le fonctionnement de l'ensemble des champs proposés pour effectuer des requêtes, notons que l'on peut effectur une recherche sur les termes médicaux d'une part mais aussi que l'on peut rechercher sur des métadonnées [34] temporelles d'autre part afin de réponre à ce type de requête : *"déterminer les prises en charges hospitalières pour la grippe d'un durée de 10[jours]"*. Pendant la saisie, un mécanisme d'autocomplétion guide l'utilisateur dans le choix des codes qu'il désire chercher ;

- Le 2e est très similaire au premier, mais il est dédié à la recherche relative aux examens biologiques et à leurs résultats. Il permet donc de retrouver l'ensemble des analyses biologiques réalisées, selon leurs paramètres et leur résultats. Ces derniers pourront être exprimés de plusieurs manières : normal/anormal, supérieur/inférieur à une valeur absolue ou positif/négatif. Le même mécanisme d'auto-complétion permettra de faciliter la saisie des termes ;

- Le 3e permet d'effectuer des recherches complexes [35] avec l'ensemble des données structurées, et de leurs métadonnées, du DPI de Rouen (pathologies, actes et biologies) (voir Figure 5.11). Cet onglet regroupe les fonctionnalités des deux précédents volets, plus la possibilité d'ordonner les éléments cherchés grâce à des opérateurs

33. Pour le moment, les termes d'indexation sont limités aux classifications CIM10 et CCAM
34. Utilisation des opérateurs booléens
35. Contraintes temporelles relatives

booléens, afin de répondre à ce type de requête : *"Déterminer l'ECG qui a été réalisé juste après une cholécystectomie"*. Les recherches complexes ne nous semblent intéressantes à réaliser au sein d'un dossier médical que dans les cas des patients ayant des histoires compliquées et suivis depuis longtemps [36]. Les résultats des recherches sont renvoyés sous forme de liste d'actes ou de séjours ainsi avec des liens pointant vers les comptes-rendus correspondants (Figure 5.12).

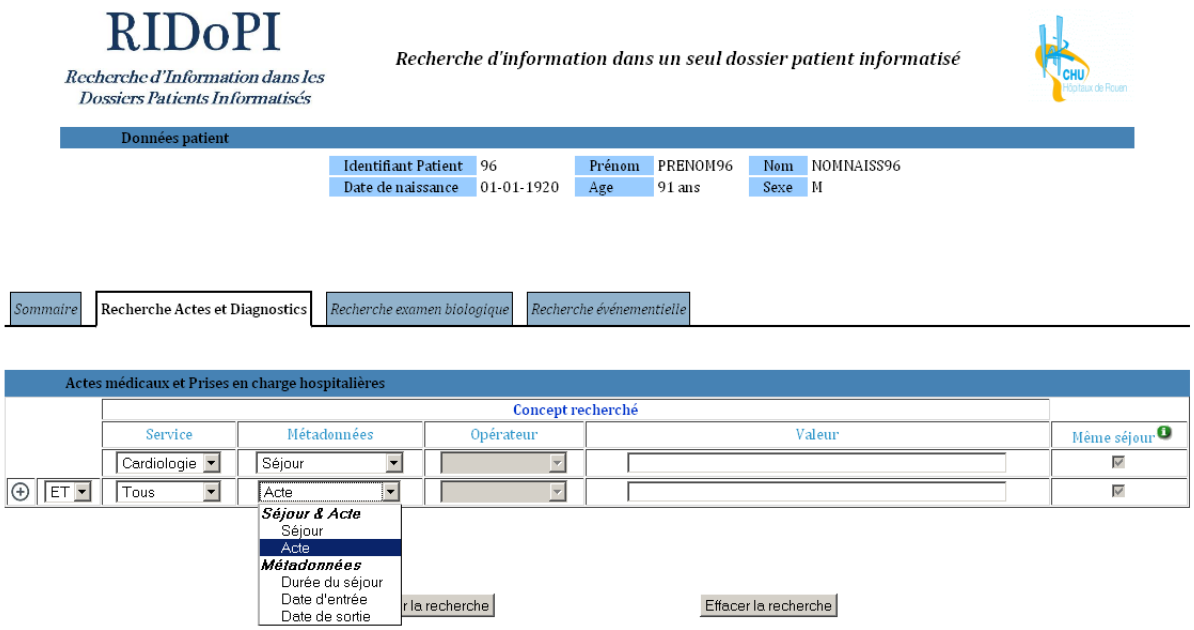

FIGURE 5.10 – Onglet "Séjour et Acte"

5.3.1.2.2 L'interface RI multi patient

L'interface de recherche multi-patient est constituée de deux champs d'interrogations dédiés, l'un à la recherche d'actes, l'autre à la recherche de diagnostics (voir Figure **??**). La recherche peut, au sein de chaque champ, s'effectuer de deux manières différentes :

1. Soit on saisit un code (CCAM ou CIM10) auquel cas l'outil renvoie l'ensemble des patients dont un acte, ou un séjour, aura été indexé par ce code ;
2. Soit on effectue une saisie en texte libre auquel cas l'outil RIDoPI retrouve tous les codes (CCAM ou CIM10) contenant ces mots et lance une recherche sur ces codes.

36. Au CHU de Roue, près de 20% des dossiers médicaux contenus dans la base CDP contiennent plus de 50 prises en charge

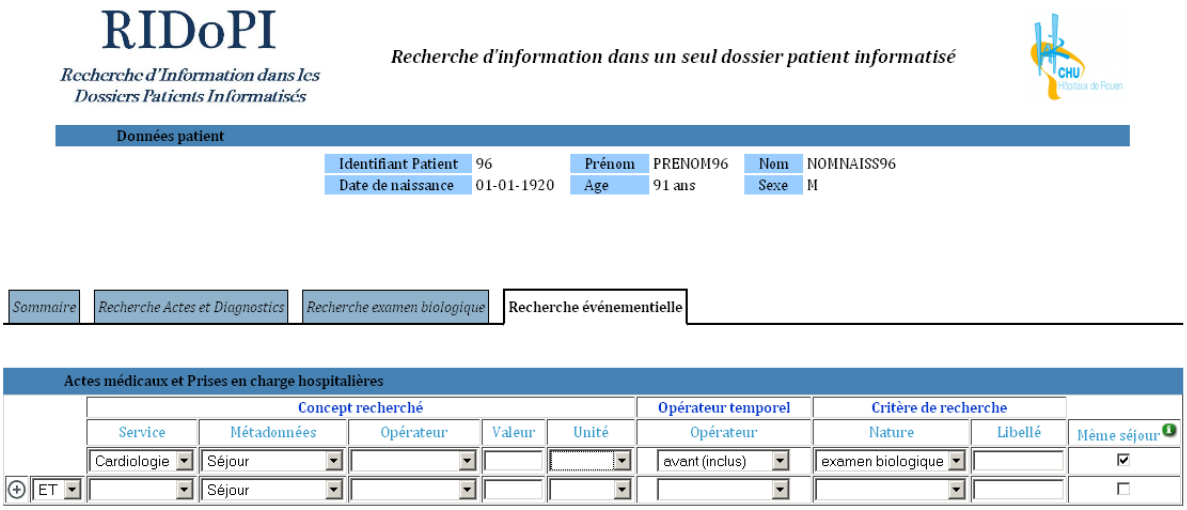

FIGURE 5.11 – Onglet "recherche événementielle"

Le mécanisme d'auto-complétion est toujours là pour guider l'utilisateur.

Les possibilités offertes par cet outil sont encore très limitées en terme d'édition de requête, toutefois, il est prévu d'y intégrer la plupart des subtilités déjà intégrées dans la version *"mono patient"* : possibilité de mettre plusieurs conditions, sur les actes et les séjours, navigation au sein des terminologies d'indexations, contraintes temporelles absolues ou relatives.

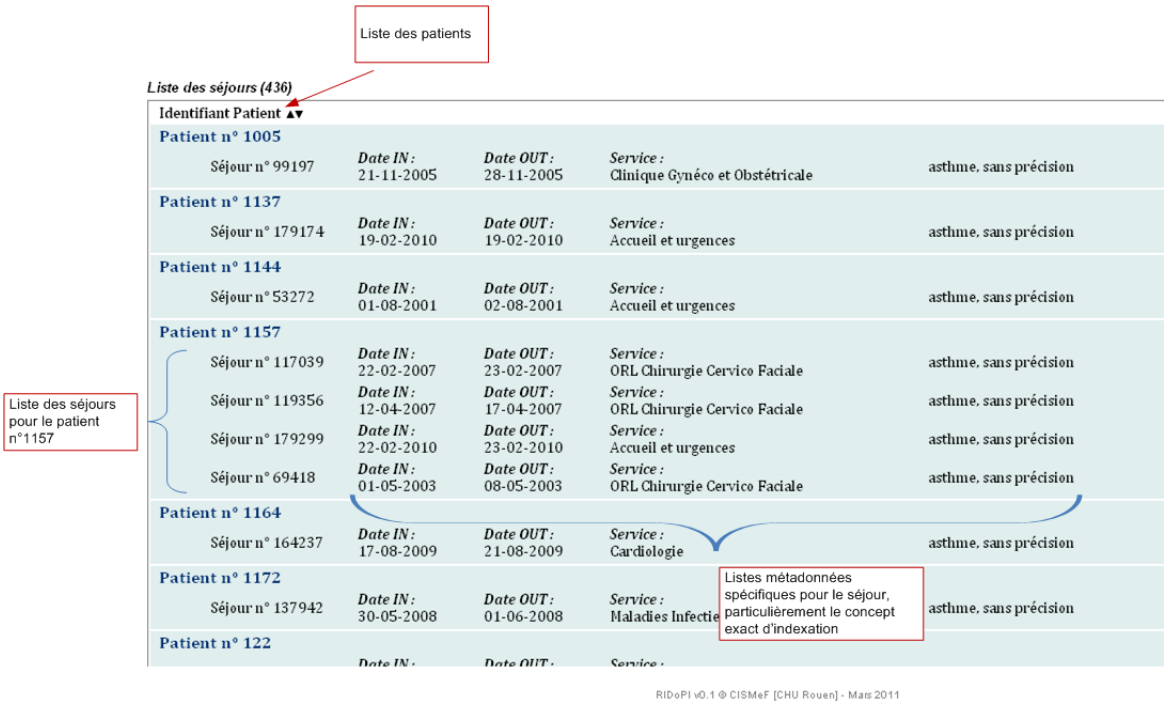

FIGURE 5.12 – Résultats des séjours diagnostiqués pour un asthme

Synthèse

Dans ce chapitre, nous avons expliqué les architectures techniques qui implémentent le modèle EI@DM proposé. La première architecture concernait l'utilisation des outils technologiques du Web Sémantique pour une RI sémantique. Cette implémentation a consisté à créer une base de données sémantique sous Oracle 11gR1, stockant les données terminolgiques en format RDF et les donneés du patient en format relationnelle. La seconde architecture n'est autre que le prototype **RIDoPI** implémenté dans le SI CISMeF. Ce prototype offre des interfaces de RI au sein d'un DPI ou d'une base de dossiers médicaux. Les deux scénarios d'implémentation nous ont permis d'expérimenter notre modèle de données générique afin de définir une structure d'évaluation. Dans le chapitre suivant, nous détaillons la validation applicative de ces deux implémentations à travers des jeux de données patients et des cas cliniques.

CHAPITRE 6

Évaluation et Résultats

Introduction

Dans les chapitre précédents, nous avons présenté notre modèle de données générique adapté à la RI avec deux scénarios d'implémentation du modèle. Les résultats ne seront pas évalués selon les critères d'évaluation spécifiques aux SRI Hersh [2008] mais seront évalués selon deux points de vue : *l'adaptation du modèle à la RI au sein d'un DPI ou d'une base de DPI* et *la complexité des requêtes*. En ce qui concerne la complexité des requêtes, nous allons étudier deux critères importants : le nombre de jointure et l'expansion sémantique. L'adaptation du modèle à la RI consite à évaluer la capacité du modèle à répondre aux besoins informationnels des professionnels. Ce critère sera analysé différemment pour les 2 scénarios d'implémentation :

- une évaluation de la pertinence des résultats pour notre base de données sémantique ;
- une évaluation de l'utilisabilité de la RI à travers le prototype **RIDoPI** pour l'accès à l'information recherchée par les professionnels.

Notre approche globale d'évaluation est résumée dans la figure 6.1.

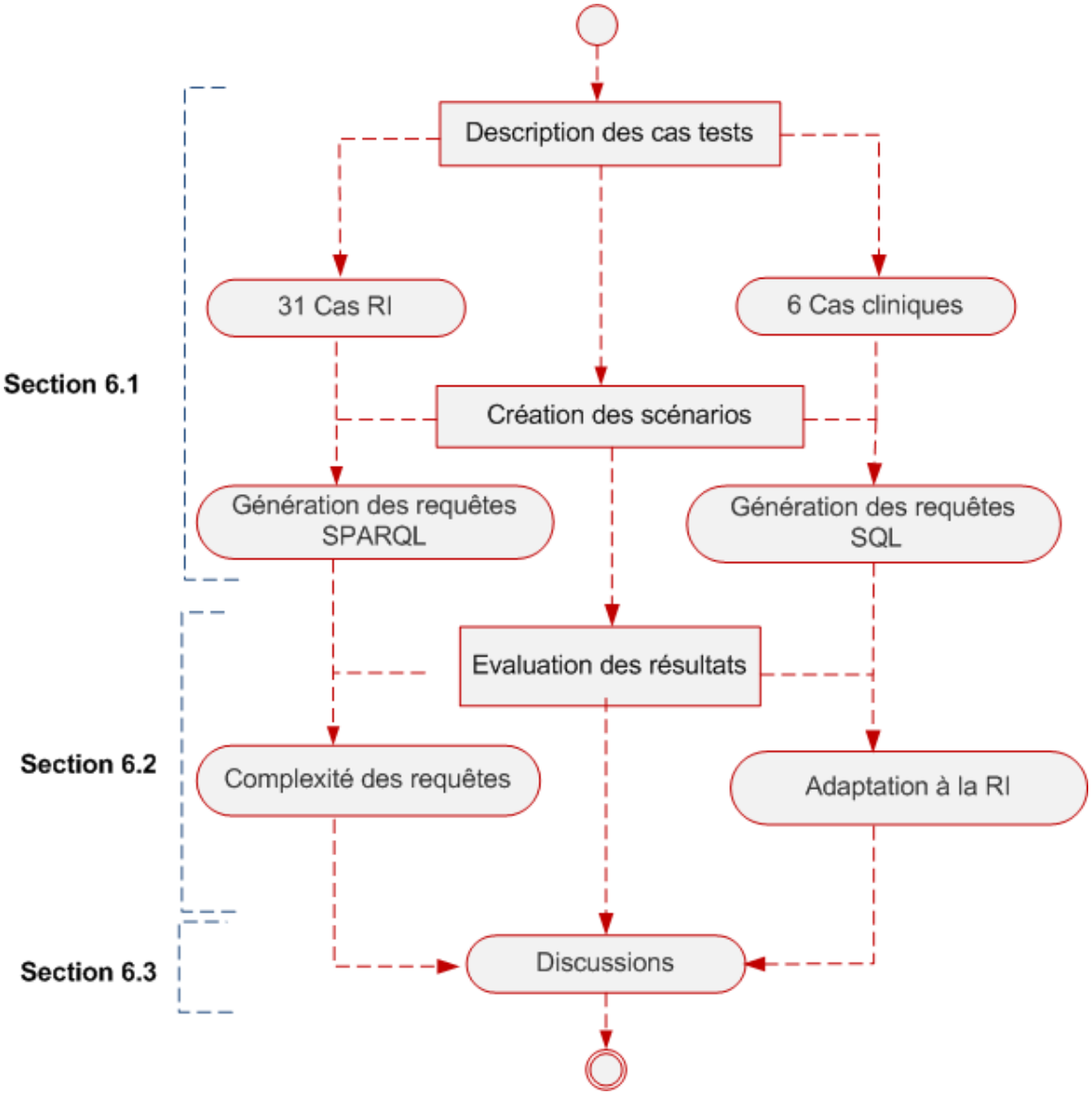

FIGURE 6.1 – Schéma synoptique de notre évaluation

6.1 Méthodologies

6.1.1 Cas tests

6.1.1.1 Cas tests pour notre base de données sémantique

Pour la préparation de ces cas tests, notre expert (PM) a tout d'abord analysé les 20 dossiers patients sélectionnés, dans leur intégralité, pour mieux comprendre la complexité médicale et la chronologie des événements médicaux afin d'en déduire une fiche résumé pour chaque dossier patient (Identifiant Patient, Age, Sexe, Nombre de pathologies codées, les pathologies exploitables mais non codées, ...). Ensuite, l'expert a défini des questions cliniques pour chaque dossier patient (au total 31 cas tests) correspondant à des besoins informationnels pertinents et utils lors d'une prise en charge d'un patient. Ces besoins informationnels pouvaient concerner des informations spécifiques sur un patient **(RI mono patient)** ou concerner des informations épidémiologiques sur plusieurs patients **(RI multi patient)**. Les questions cliniques proposées par l'expert sont constituées d'un ou plusieurs de ces 3 éléments[1] : un élément temporel ('dernière', 'même hospitalisation', '10 jours précédents', etc.), un élément alphanumérique ('TQ', '20%', etc.) et un élément concept ('radiographie', 'prostatite', 'infectieux', etc.). L'expert a exécuté manuellement ces questions cliniques sous forme de requêtes SQL dans la base CDP et a sélectionné une liste de réponses pertinentes à ces requêtes. Notre évaluation consiste à exécuter ces questions cliniques dans notre environnement sémantique et à comparer nos résultats à celles de l'expert. Ces RI[2] avaient des complexités variables, incluant l'aspect temporel entre les éléments du dossier patient. Des exemples de ces questions cliniques sont présentés dans ce tableau 6.1.

6.1.1.2 Cas tests pour notre prototype RIDoPI

Les cas tests de **RIDoPI** ont fait l'objet d'une approche différente pour leur conception. Les experts (PM et NG) ont conçu des cas cliniques sans à faire d'étude approfondie des 2000 DPI intégrés dans **RIDoPI**. Ces cas cliniques correspondent à des besoins d'informations spécifiques pouvant simplifier, par exemple le travail de recrutement des professionels pour les essais cliniques (cf. **Cas clinique 1**) ou améliorer le processus de prise en charge du patient en aidant les professionnels dans leur RI (cf. **Cas clinique 2**) :

- **Cas clinique 1** : *"Un pneumologue suspecte que la consommation d'un additif alimentaire EXXX induirait une augmentation du risque de développer un asthme. Une brève revue de la litérature l'a conforté dans son idée sans qu'il n'ait pu voir*

1. Nous excluons les opérateurs booléens (ET, OU, ...), arithmétiques (<, >, ...) ou qualificatifs,...
2. Voir en Annexe A.1.2 des exemples de fiche résumé de DPI et les questions cliniques associées

Question clinique	**Type de RI**	**Terme(s)** [3] **de la RI**
Q1 : Rechercher la dernière radiographie du patient X	RI mono patient sur un acte	(radiographe, ECG,...)
Q2 : Quelles sont les épisodes infectieux du patient Y ?	RI mono patient sur des séjours	(infectieux, infection, ...)
Q3 : Rechercher les patients pour lesquels sont codés à la fois prostatite et SIDA dans un même séjour	RI multi patient sur des séjours avec plusieurs diagnostiques	(prostatite, SIDA, VIH, AID, ...)
Q4 : Rechercher s'il y a eu des complications des anticoagulants et s'il y a eu un TQ $< 20\%$ dans les 10 jours précédents la date de codage	RI mono patient et chronologique sur des examens biologiques associés à des diagnostiques	(anticoagulant, TQ, ...)
Q5 : Rechercher tous les patients avec un infarctus du myocarde et une bilirubine supérieure à 20 et pendant la même hospitalisation	RI multi patient et chronologique sur des examens biologiques associés à des diagnostiques	(infarctus myocarde, bilirubine, ...)

TABLE 6.1 – Description des questions cliniques

que la relation avait déjà été mise en évidence chez l'homme. Il projète donc de réaliser une étude cas-témoin pour répondre à cette question. Les calculs de nombre de sujet nécessaire lui ont révélé qu'il devait inclure 185 cas et 370 témoins dans son étude... Il cherche donc dans l'outil RIDoPI les patients atteints d'asthme pris en charge au CHU" ;

– **Cas clinique 2** :*"Monsieur T, est hospitalisé en urgence pour des douleurs abdominales. L'interrogatoire révèle un antécédent de chirurgie sur l'aorte (le patient ne se souvient plus très bien, mais nous montre une cicatrice de laparotomie xyphopubienne), réalisée dans ce même CHU quelques années auparavant, et un arrêt des matières et des gaz depuis plusieurs heures. L'imagerie révèle une occlusion haute. Avant d'opérer, le chirurgien souhaiterait avoir plus d'informations sur l'opération précédente et cherche donc dans* ***RIDOPI*** *l'ancien compte-rendu opératoire de Mr T."*

L'objectif est d'évaluer la capacité du modèle à faciliter la RI dans ces différents cas cliniques [4]. Par ailleurs, nous nous sommes pas limités à ces cas cliniques pour évaluer le prototype **RIDoPI**. [Deshmukh et al., 2009] ont définis des cas cliniques pour évaluer le *Framewkork I2B2* selon plusieurs critères, nous avons évalué le prototype **RIDoPI** avec ces cas cliniques (cf. paragraphe 6.2.2.3).

4. Voir en Annexe A.1.1 quelques exemples d'autres cas cliniques.

6.1.2 Construction de scenarios

Après avoir décrit les cas tests, nous donnons quelques exemples de construction de scenarios à travers l'exécution des cas tests dans les environnements informatiques déployés pour les deux scenarios d'implémentation de notre modèle.

6.1.2.1 De la question utilisateur à la requête SPARQL

A titre d'exemple, prenons la question clinique Q2 : *"Quelles sont les épisodes infectieux du patient Y ?"*. L'exécution de la requête dans notre base de données sémantique se fait en deux étapes (voir Tableau 6.2) :

- **Etape 1** : nous déterminons l'ensemble des concepts CIM-10 dont le métaterme est **"Infectiologie"** et nous effectuons une explosion hiérarchique sur chacun de ces concepts. Cette explosion est une solution parmi tant d'autres pour obtenir une expansion sémantique de notre requête ;
- **Etape 2** : nous déterminons tous les séjours d'hospitalisation du patient dont les diagnostics sont codés avec ces concepts CIM-10. Conformement au modèle EI@DM, la partie SQL de la requête SPARQL consiste en une jointure [à gauche] entre la table **RelationDescripteur_EI** et la table **RelationInformation_EI** pour déterminer les EIs liés entre eux par une relation de type *"Patient nº 1 - Sejour"* correspondant à la métadonnées *T_REL_SEJ_PAT* ; ces EIs étant indexés avec les concepts CIM10 issues de l'**Etape 1**.

La requête figurant dans le tableau 6.2 montre que les deux étapes sont exécutées simultanément grâce aux opérateurs sémantiques d'Oracle (SEM_MATCH) qui permettent d'interroger en même temps des données relationnelles et des données triplets.

6.1.2.2 De la question utilisateur à la requête SQL

Nous prenons, comme exemple, le **Cas clinique 2**. Le tableau 6.3 décrit la requête RI exécutée à travers le prototype **RIDoPI**. Pour cette question clinique, nous déterminons tous les séjours du patient Mr T. ainsi que les CRs d'hospitalisation. Conformément au méta modèle CISMeF, cette requête SQL détermine d'abord tous les séjours `S` du patient Mr T. dans la table **TB_OBJECT_PROPERTY** avec la contrainte *(DM_IS_STAY)*[5] ; ainsi que les comptes-rendus d'hospitalisation (CRH) avec une jointure [à gauche] sur une autre table **TB_OBJECT_PROPERTY** selon la contrainte *(DM_IS_REC)*[6] ; ces séjours étant indexés par les concepts CIM10 codant toutes les

5. S *est_sejour_de* du patient Mr T.
6. CRH *est_un_compterendu_de* de S

pathologies sur l'aorte.

Question	**Requête SPARQL**
Rechercher les épisodes infectieux du patient n° 1 ?	SELECT sejour.ID_EI, patient.ID_EI1, DATE_RELATION FROM RelationDescripteur_EI sejour **LEFT OUTER JOIN** RelationInformation_EI patient ON sejour.ID_EI = patient.ID_EI2 WHERE patient.ID_TypeRelation= 'T_REL_SEJ_PAT' and patient.ID_EI1= 'PATIENT1' and sejour.ID_DESC_URI IN ***(SELECT fils FROM TABLE (SEM_MATCH('*** ***(?r <rdf :type> <publishing#BT-NT>)*** ***(?r <publishing#BT> ?pere)*** ***(?r <publishing#NT> ?fils)*** ***(?pere <smts#CIM10MT> "infectiologie")',*** ***SEM_MODELS('CIM10'),null,null,null)))***

TABLE 6.2 – Requête SPARQL dans notre base sémantique

Question	**Extrait Requête SQL**
Cas clinique 2	SELECT DISTINCT sej.RDF_RESOURCE_SOURCE, crsej.RDF_RESOURCE_SOURCE, sejindex.RDF_RESOURCE_INDEX FROM TB_INDEXING sejIndex **LEFT OUTER JOIN** TB_OBJECT_PROPERTY crsej ON crsej.RDF_RESOURCE_TARGET = sejIndex.RDF_RESOURCE_INDEXED **LEFT OUTER JOIN** TB_OBJECT_PROPERTY sej ON sej.RDF_RESOURCE_SOURCE = sejIndex.RDF_RESOURCE_INDEXED WHERE sejIndex.TYPE_ID='DM_INDEXING' and crsej.TYPE_ID='DM_IS_REC' and sej.TYPE_ID='DM_IS_STAY' and sejIndex.RDF_RESOURCE_INDEX IN **(Liste des concepts CIM10 codant toutes les pathologies sur l'aorte)** [7] ;

TABLE 6.3 – Requête SQL à travers le prototype RIDoPI

6.2 Résultats

Nous avons évalué les résultats des 2 scenarios d'implémentation de notre modèle EI@DM selon deux points de vue : la complexité des requêtes et l'adaption du modèle à la RI.

6.2.1 Complexité des requêtes

6.2.1.1 Requête et Jointure

Les exemples de RI étudiés concernent des questions utilisateurs de complexité intermédiaire :

1. Recherche des séjours liés à une pathologie infectieuse chez le patient n° 1 ;
2. Recherche d'un acte radiologique : par exemple la radiographie du bassin chez le patient n° 6.

Nous avons analysé le nombre de jointures des requêtes uniquement sur notre base de données sémantique. En effet, du fait de l'équivalence entre notre modèle de données et le méta-modèle CISMeF, nous avons le même constat pour la partie SQL des requêtes SPARQL [sur notre base de données sémantique] que sur les requêtes SQL simples [sur la base de données du prototype **RIDoPI**], à savoir UNE JOINTURE commune à l'ensemble des requêtes.
Le tableau 6.4 compare les requêtes SQL simples exécutées dans CDP et les requêtes SQL, exécutées dans notre base sémantique de données, intégrant donc des graphes SPARQL. les requêtes dans la base CDP contiennent 4 tables et 3 jointures pour chacune d'entre elles, mais les tables et les jointures changent selon la donnée qu'on recherche. Les requêtes SQL, sur le modèle EI@DM, n'incluent que 2 tables et une jointure. Cette dernière est commune à l'ensemble de nos requêtes car la table **RelationInformaion_EI** contient l'ensemble des EIs et leurs relations entre eux. Ainsi dans le modèle que nous proposons, la RI peut être réalisée via des requêtes types dont le seul élément variable est la partie terminologique de la clause *"WHERE"*.

6.2.1.2 Requête et Expansion sémantique

Les concepts CIM10 correspondant à une pathologie infectieuse ne peuvent être déterminés directement et nous utilisons la notion de métaterme (un super-concept pour lier des concepts de même spécialité médicale [Thirion et al., 2003]) pour rechercher ces concepts. Dans les 2 scenarios d'implémentation, l'expansion sémantique est faisable

grâçe à la disponibilité de plusieurs vocabulaires médicaux [francophones] constituant l'univers multi terminologique et aux outils automatiques d'alignement de concepts développés par l'équipe CISMeF. Cette approche sémantique, en utilisant des métatermes, est une approche parmi tant d'autres pour améliorer la RI. Dans le tableau 6.4, la notion de métaterme est utilisée, à savoir le métaterme *"infectiologie"* pour déterminer les concepts CIM10 liés à cette spécialité. A la seule différence que pour la requête SPARQL, nous avons étendu ces concepts CIM10 en exploitant leurs relations hiérarchiques et cela nativement. Cette expansion sémantique, non applicable dans la base CDP, nous permet de faire une RI sémantique. Le même principe est appliqué sur les requêtes SQL dans **RIDoPI** par une explosion sur les relations hiérarchiques.

Requête CDP	**Requête SPARQL**
SELECT p.IDPATIENT, s.IDPECUF_RM, s.DATEENTREE FROM METATERMES_CIM mt **INNER JOIN** DIAG d ON mt.CODE = d.CODECIM10 **INNER JOIN** (PATIENT p **INNER JOIN** PEC s ON p.IDPATIENT = s.IDPATIENT) ON d.IDPECUF_RM = s.IDPECUF_RM WHERE (((p.IDPATIENT)=1) AND ((mt.METAT)="infectiologie"))	SELECT sejour.ID_EI, patient.ID_EI1, DATE_RELATION FROM RelationDescripteur_EI sejour **LEFT OUTER JOIN** RelationInformation_EI patient ON sejour.ID_EI = patient.ID_EI2 WHERE patient.ID_TypeRelation= 'T_REL_SEJ_PAT' and patient.ID_EI1= 'PATIENT1' and sejour.ID_DESC_URI IN (SELECT c2 FROM TABLE (SEM_MATCH(' (?r <rdf :type> <publishing#BT-NT>) (?r <publishing#BT> ?c1) (?r <publishing#NT> ?c2) (?c1 <smts#CIM10MT> "infectiologie")', SEM_MODELS('CIM10'),null,null,null)))
SELECT PATIENT.IDPATIENT, PEC.IDPECUF_RM, ACTE.DATEACTE, CCAM.LIBELLECODE FROM CCAM **INNER JOIN** (ACTE **INNER JOIN** (PATIENT **INNER JOIN** PEC ON PATIENT.IDPATIENT = PEC.IDPATIENT) ON ACTE.IDPECUF_RM = PEC.IDPECUF_RM) ON CCAM.IDACTE = ACTE.IDACTE WHERE (((PATIENT.IDPATIENT)=1) AND ((CCAM.["LIBELLECODE"]) Like "bassin*"))	SELECT acte.ID_EI, sejour.ID_E1, DATE_RELATION FROM RelationDescripteur acte **LEFT OUTER JOIN** RelationInformation sejourActe ON acte.ID_EI = sejourActe.ID_EI2 **LEFT OUTER JOIN** RelationInformation sejour on sejour.ID_EI1 = sejourActe.ID_EI1 and sejour.ID_EI2 ='PATIENT6' WHERE sejour.ID_TypeRelation = 'T_REL_SEJ_ACT' and acte.ID_DESC_URI IN (SELECT c2 FROM TABLE (SEM_MATCH(' (?c2 <rdf-schema#label> ?label)', SEM_MODELS('DM_CCAM'), null, null, null)) WHERE REGEXP_LIKE (LOWER(label),'((.)*(bassin)(.)*)'))

TABLE 6.4 – Comparaison des requêtes SQL-CDP et des requêtes SPARQL

6.2.2 Adaptation du modèle à la RI

6.2.2.1 Pertinence de la RI

Les résultats des 31 cas tests ont été manuellement évalués par l'expert du domaine et ont été classés comme suit : **"Pertinent"** quand tous les EIs sont restitués, **"Incomplet"** si certaines EIs ne sont pas restitués et **"Non pertinent"** si aucuns EIs ne sont restitués ou qu'ils ne correspondent pas aux EI attendus. Sur les 31 cas tests , 25 tests cases sont spécifiques à un patient donné et 6 cas tests pour l'ensemble de la base des 20 DPI.
Sur les 25 *requêtes mono patient*, 9 correspondaient à des recherche d'un ou plusieurs séjours liés à un (ou plusieurs) diagnostic(s), 6 à la recherche d'un (ou plusieurs) acte(s), 2 à la recheche de prises en charges dans un service donné, 5 pour la recherche d'évènements cliniques [8] et 3 à la survenue de types de pathologies (par exemple un accident médicamenteux ou un épisode infectieux).
Pour les 6 *requêtes multi patient* , 2 recherches concernent des données démographiques et des pathologies, 3 des types de pathologies ou d'actes (par exemple une intervention chirurgicale) et 1 des évènements cliniques.
Les résultats de cette évaluation sont résumés dans les tableaux 6.5 et 6.6.
Pour 17 cas des 25 *requêtes mono patient* (68% ; IC95% = [46.5%-85.1%]) et pour 5 cas des 6 *requêtes multi patient* (83% ; IC95% = [35.9%-99.6%]), les résultats des recherches étaient conformes aux résultats attendus. Les autres résultats étaient soit incomplets, soit faux. L'analyse de ces erreurs a permis de mettre en évidence trois types d'erreurs : le tableau 6.7 résume les types d'erreurs et niveau de difficulté à les résoudre.

	Conformité des résultats			
Types de requêtes	n	**Pertinent**	**Incomplet**	**Non pertinent**
Diagnostics	9	7	2	0
Actes	6	5	1	0
Prises en charge	2	0	0	2
Aspectw temporels	5	5	0	0
Types de pathologies	3	0	0	3

TABLE 6.5 – Tableau d'évaluation des résutats pour la RI mono patient

8. Prise en compte de l'aspect temporelle

Types de requêtes	Conformité des résultats			
	n	**Pertinent**	**Incomplet**	**Non pertinent**
Données diagnostiques et données démographiques	2	2	0	0
Types de pathogies ou actes	3	2	0	1
Aspects temporels	1	1	0	0

TABLE 6.6 – Tableau d'évaluation des résultats pour la RI multi patient

Requête utilisateur	Type Interprétation	Difficulté
Quelles sont les épisodes infectieux du patient Y ?	Regroupement de concepts (maladies infectieuses)	++
Rechercher les séjours pendant lesquels il y a eu une pneumopathie	Plusieurs synonymes (pneumopathies, pneumonie, pneumologie, pneumocoque, ...)	+
Rechercher une IRM ou un scanner du crane	Problème de Terminologies [d'interface] (imageries célébrales)	+
Rechercher quel séjour est du à une complication d'un kyste de l'ovaire	Informations non indexées (dans les antécédents des CRs)	+++
Lister les accidents médicamenteux	Type de pathologie, Concept inexistant	+++
Rechercher des prises en charge aux urgences	Erreur Codage PMSI *(*)* ou Information non codée	+
Rechercher s'il y a eu des complications des anticoagulants	Concept complexe	+++
Rechercher une intervention chirurgicale	Type d'acte médical, 1er caractère (A ou C) d'un code CCAM	++
Y-a-t-il eu une hémorragie digestive, si oui et quand ?	Type de pathologie, Concept inexistant	++

TABLE 6.7 – Types d'erreurs et Niveau de difficulté

6.2.2.2 Fonctionnalités et utilisation des interfaces de RI

L'évaluation du prototype **RIDoPI** n'a porté que sur l'*utilisabilité*[9] de la RI. Les fonctionnalités mises en place au sein des interfaces de RI ont permis de formuler tous les cas cliniques ainsi que, certains cas cliniques issues de l'évaluation du *framework I2B2* (cf. paragraphe 6.2.2.3). Prenons l'exemple du **Cas clinique 2**, d'un point de vue IHM, l'utilisateur (le chirurgien) va chercher toute pathologie [ou tout acte] sur l'aorte en utilisant comme mot clé de recherche *"aort"* dans l'interface **RI mono patient** (voir Figure 6.3) à travers la requête *booléenne* : ***"aort*[CIM 10] OR aort*[CCAM]"***. L'interprétation de la requête est la suivante : " rechercher tous les séjours ou les actes dont les codes CIM 10 ou les codes CCAM contiennent le terme "aort". Rapidement, il peut accéder directement aux différents comptes-rendus d'hospitalisation ou d'actes qui l'intéresse (voir Figure 6.4). L'aspect de la multi terminologie est considéré par le fait que l'utilisateur va rechercher dans les deux terminologies : la CIM 10 pour les pathologies et la CCAM pour les actes.
Pour les cas cliniques de [Deshmukh et al., 2009], nous prenons, comme exemple d'évaluation, la requête n° 9 : *"PI needs to find d-dimers values, dates and times on patients who have undergone neurosurgery"*[10]. La figure 6.2 représente l'interface d'exécution de cette requête.

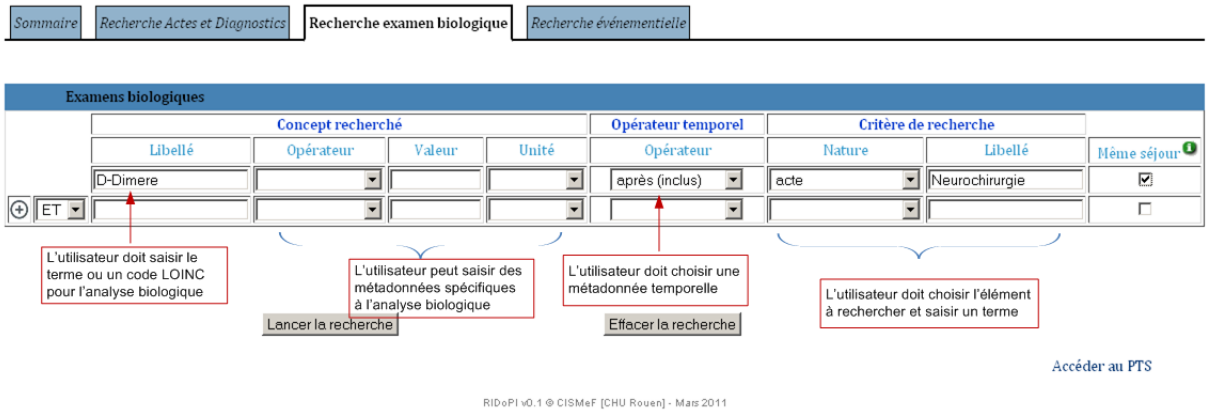

FIGURE 6.2 – Requête n° 9 : "Liste des d-dimères des patients ayant subi une intervention neurochirurgicale", exécuté dans le prototype **RIDoPI**

9. L'*utilisabilité* ou *usabilité* est définie par la norme ISO 9241-11 comme « le degré selon lequel un produit peut être utilisé, par des utilisateurs identifiés, pour atteindre des buts définis avec efficacité, efficience et satisfaction, dans un contexte d'utilisation spécifié » [Définition WIKIPEDIA]

10. Le médecin veut la liste des patients qui ont eu des dosages d-dimères et une opération neurochirurgicale pendant le même séjour

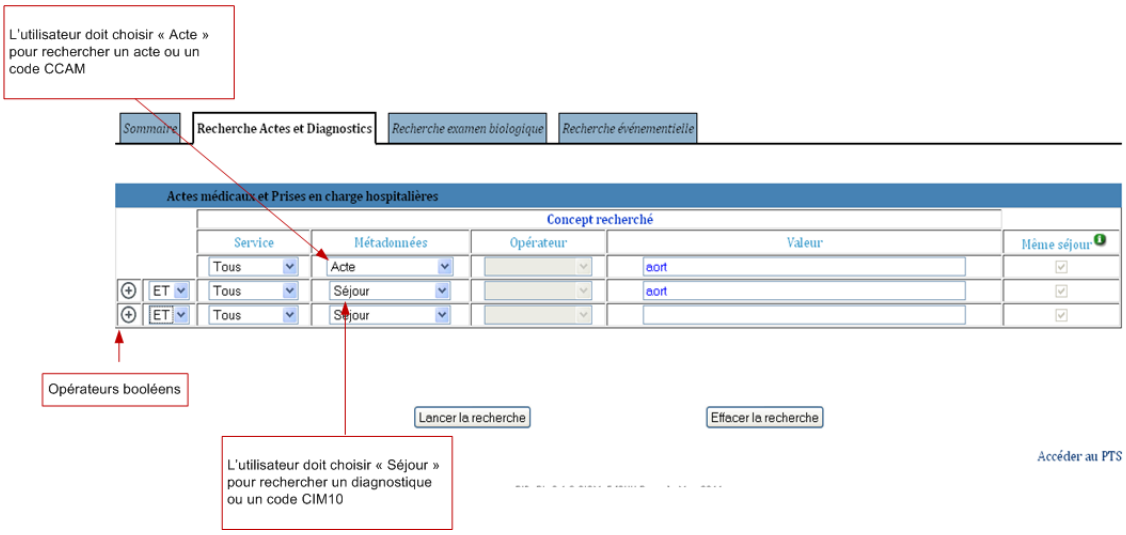

FIGURE 6.3 – Requête utilisateur dans **RIDoPI** pour le **Cas clinique 2**

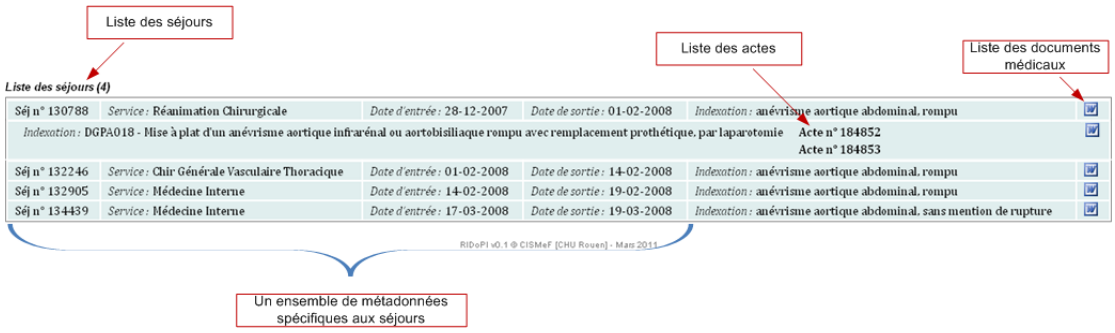

FIGURE 6.4 – Résultats de la requête utilisateur dans **RIDoPI** pour le **Cas clinique 2**

6.2.2.3 Conformité entre l'outil I2B2 et le prototype RIDoPI

Nous avons réalisé une analyse comparative entre **I2B2** et **RIDoPI** pour évaluer leur conformité fonctionnelle. Nous utilisons, comme méthode d'évaluation, celle implémentée, par [Deshmukh et al., 2009] pour évaluer **I2B2**. Cette méthode incluait 27 requêtes sélectionnées [11] et évaluées à travers 9 critères répartis classifiant ces requêtes en 2 types :

- des requêtes incluant des traitements (***avec*T**), parmi lesquelles :
  - + des traitements spécifiques (***spec*T**) [12] ;
  - + des pré(post) traitements (***pré(post)*T**) [13] ;

11. Voir en Annexe A.2 Les 27 requêtes, les types de données prient en compte par ces requêtes ainsi que les résultats de l'évaluation
12. Des traitements propriétaires à **I2B2**
13. Des traitements en amont sur les données et les questions utilisateurs, des traitements en aval sur

+ des calculs de champs (***calc*T**) [14] ;
+ des opérations d'ajouts de nouveaux attributs (***newAtt*T**) [15] ;
+ des opérations de modifications de métadonnées existantes (***mt*T**) [16] ;
+ des conditions d'exception (***excep*T**) [17] ;
+ des conditions temporelles (***temp*T**) [18] ;

– des requêtes pouvant être exécutées sans traitement (***sans*T**) directement à travers les interface de l'outil **I2B2** ;

Nous avons repris le tableau de [Deshmukh et al., 2009] dans lequel nous avons ajouté les colonnes grisées correspondant à notre évaluation des 27 requêtes sur le prototype **RIDoPI**. Les lignes jaunes correspondent aux requêtes exécutées dans **I2B2** sans traitement. Nous avons calculé, pour chaque critère, les proportions de requêtes les utilisant. Certains critères de [Deshmukh et al., 2009] sont ambigus comme les conditions d'exception et donc nous n'avons pas fait de comparaison sur ces critères(0% pour **RIDoPI**).
Pour comprendre l'évaluation, nous prenons deux requêtes comme démonstration :
A titre d'exemple, la requête n° 4 : *"PI would like a list of patients who had hip/knee surgery and had received Rifampicin after their surgery"* [19]. Sont inclus ces types de données : un médicament *"Rifampicin"*, une procédure *"hip/knee surgery"* et une contrainte temporelle [relative] ou plus précisement chronologique *"after"*. Pour **I2B2**, cette requête implique un post-traitement et la gestion de la contrainte temporelle. Pour **RIDoPI**, nous pouvons l'exécuter directement sur les interfaces en ajoutant les informations relatives aux prescriptions médicamenteuses contenues dans les courriers et comptes-rendus médicaux, sous forme d'un type d'actes particulier.
Autre exemple, la requête n° 17 : *"This is a pre-research data request for an estimate of sample size on patients with non-ruptured cerebral aneurysm (ICD Code 437.30) who have had at least two Head MRA procedures"* [20]. Elle inclus un diagnostic *"non-ruptured cerebral aneurysm"*, un acte *"Head MRA"* et une contrainte sur le nombre d'occurences (*at least tow*). **I2B2** y répond facilement alors que pour **RIDoPI**, cette requête implique un traitement de plus sur la gestion des occurences.
Le tableau 6.8 résume les résultats de l'évaluation.

---

les résultats des requêtes machines, d'autres traitements spécifiques, ...

14. Des traitements sur l'agrégation des données
15. Traitement nécéssitant l'ajout d'une nouvelle donnée
16. Des traitements impliquant la modification de métadonnées existantes...
17. Traitement de cas d'exception...
18. Besoins d'opérateurs temporelles relatives tels que ceux-là (*"même séjour, avant, après, durant, ..."*)
19. L'utilisateur veut la liste des patients qui ont reçu de la Rifampicine après avoir bénéficiés d'une chirurgie de hanche ou de genou
20. L'utilisateur veut une estimation du nombre de patient ayant un anévrysme cérébral qui ont bénéficié d'au moins deux Angio-IRM cérébrales

n° Req	*sans*T		*avec*T															
			*spec*T		*pré*T		*post*T		*excep*T		*temps*T		*calc*T		*newAtt*T		*mt*T	
1					*	*	*	*				*			*	*	*	
2	*	*				*					*							
3					*		*	*					*	*	*	*	*	
4		*					*				*					*		
5	*				*				*			*				*		
6	*	*																
7					*		*	*					*	*		*	*	
8		*			*		*				*	*	*				*	
9	*	*		*			*											
10	*			*		*	*									*		
11					*	*	*	*			*		*	*	*		*	
12		*			*										*		*	
13			*		*		*	*			*		*		*	*	*	*
14		*	*		*		*	*						*	*		*	
15	*	*														*		
16	*	*		*														
17	*			*				*						*				
18		*			*			*					*		*	*	*	
19	*	*									*					*		
20			*		*			*			*			*	*	*	*	
21(+)			*		*	*		*						*	*	*	*	*
22(+)			*		*	*	*	*	*		*		*	*	*	*	*	*
23(+)			*		*	*	*	*	*		*			*	*	*	*	*
24	*	*						*			*							
25			*		*	*	*	*			*		*		*	*	*	*
26	*	*					*	*			*		*	*				
27	*	*		*			*				*							
Total	12	14	7	5	14	7	15	15	3	0	13	3	9	10	12	15	14	5
%	44%	52%	26%	19	52%	26%	55%	55%	11%	0%	48%	11%	33%	37%	44%	56%	52%	19%

TABLE 6.8 – Évaluation I2B2 et RIDoPI

6.2.2.3.1 Synthèse de la conformité fonctionnelle

Globalement, nous n'avons pas mis en évidence de différences significatives entre les deux outils concernant le nombre de requêtes pouvant être exécutées sans traitement (44% vs 52%). Depuis l'évaluation de [Deshmukh et al., 2009], **I2B2** a évolué et s'est sans doute améliorée. L'intégration de données tels que les prescriptions médicamenteuses, des données gérées par le SIC et non contenues par le DPI [21] et d'autres métadonnées contenues dans les courrier et les comptes-rendus médicaux permettront de répondre à plus de requêtes.
Cependant, nous avons observé des résultats significatifs [22] uniquement pour deux critères : le traitement de la temporalité et les modifications dans les métadonnées prédéfinies (respectivement 11% et 19% pour **RIDoPI** et 48% et 52% pour **I2B2**). Pour le premier critère, les interfaces d'**I2B2** ne sont pas adaptées pour formuler une requête utilisateur contenant des contraintes chronologiques relatives (la requêten° 12 : *"List of patients who have had D-dimer tests ordered or duplex ultrasound studies performed, along with the test/procedure dates and test-results"* [23]). **RIDoPI** permet, quant à lui, de formuler des requêtes événementielles dans 90% des cas. Les interfaces de RI ne sont pas, toutefois, simples à prendre en main pour l'utilisateur.
Nous ne sommes pas comparables sur certaines requêtes avec l'outil **I2B2**, par le seul fait que c'est un outil dédié pour exploiter les données à des fins d'analyses statistiques. La majorité des post-traitements, pour **RIDoPI**, concerne la gestion des formats de sortie des résultats. Notre objectif n'est pas de faire des statistiques sur nos données mais de rechercher une information et la présenter sous une forme adaptée à l'utilisateur. Ce dernier pourra, donc, les exporter vers d'autres outils de calculs pour les exploiter.
Les requêtesn° 21, 22 et 23 n'ont pas été simples à évaluer sur notre prototype car elles impliquent l'intégration de sources externes de données, qui pourrait nous obliger à adapter nos parseurs d'intégration et/ou d'ajouter et de définir de nouvelles métadonnées dans notre modèle. C'est quasiment la seule situation qui impose de modifier les métadonnées. Le modèle EI@DM est flexible, ce n'est pas le cas pour **I2B2** qui, dans 51% des cas de ces requêtes impliquent des modifications dans ce framework. Ce dernier, néanmoins, par sa modularité, permet d'intégrer des nouveaux modules (*hive*) pour gérer ses modifications tels que les travaux de [Mate et al., 2011]...

21. les référentiels sur les clinicens prenant en charge le patient, certaines données démographiques, ...
22. Comparaison à l'aide d'un tests Chi2
23.

6.3 Discussions sur cette évaluation

6.3.1 Sémantique et Syntaxe des requêtes

L'approche de mise en oeuvre de cette expansion est différente syntaxiquement dans une requête SQL incluant des graphes SPARQL et une requête SQL simple mais le principe reste le même. L'utilisation de requêtes incluant des graphes SQL et l'expansion sémantique doivent être améliorés par des algorithmes de post-traitement et des techniques d'optimisation pour diminuer le nombre de jointure des requêtes exécutées et réduire la complexité des graphes SPARQL dans les 2 scenarios d'implémentation de notre modèle EI@DM. Ceux-ci feront l'object d'une étude approfondie en guise de perspective, pour évaluer la *scalabité* du langage SPARQL et fournir un *preuve de concept* sur l'apport des technologies sémantiques à la RI.

6.3.2 La recherche d'information multi terminologique (RIMT)

Les vocabualaires controlés implémentés intégralement, à savoir les classifications CIM10 et CCAM , la nomenclature LOINC, ont permis d'évaluer l'adaptation du modèle proposé à la RI. Dans le cadre des deux scénarios d'implémentation de notre modèle EI@DM, la RIMT en tant que telle et son apport pour une RI sémantique n'ont pas été évalués. Les six résultats *Incomplets* et *Non pertinents* des tableaux 6.5 et 6.6 sont dues à une mauvaise interprétation de la question utilisateur ; des erreurs qui auraient donc pu être évitées. Rappelons que la validation des termes de cette requête a été faite manuellement. Cette étape d'interprétation est centrale dans la RI fondée sur des terminologies/ontologies médicales. Plusieurs méthodes ont été implémentées dans le moteur de recherche **Doc'CISMeF** pour pallier à ces différentes erreurs, et ont montré leur efficacité, parmi lesquelles :

1. La création de synonymes au sein des terminologies qui permet de limiter les situations où une recherche n'aboutit pas, par plusieurs expressions désignant une même notion (échocardiographie et échographie cardiaque, pneumonie franche lobaire aiguë et pneumopathie à pneumocoque) [Douyère et al., 2004] ;

2. Les métatermes qui sont utilisés pour regrouper plusieurs concepts de différents terminologies dans une même spécialités [Gehanno et al., 2007; Massari et al., 2008] ;

3. La racinisation et la lemmatisation des termes des requêtes utilisateurs, des techniques qui permettent de s'affranchir des variations flexionnelles des mots en les

résumant à leur racine ou à leur lemme. Grâce à elles, l'utilisateur cherchant les "douleurs du thorax" trouvera la "douleur thoracique" [Soualmia et al., 2006] ;

4. Les stratégies de RI qui correspondent à des requêtes préparées pour des mots clés souvent employées par les utilisateurs mais n'existant pas ou existant sous une autre forme au sein des terminologies [Gehanno et al., 2007].

Le tableau 6.9 complète le tableau 6.7 en proposant les solutions pour améliorer, dans le prototype **RIDoPI**, l'interprétation des questions utilisateurs.

Requête utilisateur	Type Interprétation	Solution possible
Quelles sont les épisodes infectieux du patient Y ?	Regroupement de concepts	Métatermes, Stratégies de RI
Rechercher les séjours pendant lesquels il y a eu une pneumopathie	Plusieurs synonymes (pneumopathies, pneumonie, pneumologie, pneumocoque, ...)	Synonymie, Lemmatisation/Stemming
Rechercher une IRM ou un scanner du crane	Problème de Terminologies [d'interface] (imageries célébrales)	Synonymie, Lemmatisation/Stemming
Rechercher quel séjour est du à une complication d'un kyste de l'ovaire	Informations non indexées (dans les antécédents des CRs)	Indexation sémantique, Extracion de valeurs et/ou d'états, ...
Lister les accidents médicamenteux	Type de pathologie, Concept inexistant	Stratégies de RI
Rechercher des prises en charge aux urgences	Erreur Codage PMSI *(*)* ou Information non codée	Métatermes, Stratégies de RI
Rechercher s'il y a eu des complications des anticoagulants	Concept complexe	Stratégies de RI
Rechercher une intervention chirurgicale	Type d'acte médical, 1er caractère (A ou C) d'un code CCAM	Stratégie de RI, Métatermes, ...
Y-a-t-il eu une hémorragie disgestive, si oui et quand ?	Type de pathologie, Concept inexistant	SNOMED

TABLE 6.9 – Types d'erreurs et Solutions

6.3.2.1 Indexation des données non structurées

La limite de notre travail est due à l'utilisation, pour la RI, des données structurées nativement et seulement celles-ci. Il existent deux solutions : (a) augmenter la structuration des données en favorisant l'indexation à la source et en mettant à disposition des professionnels des outils d'aide au codage [Musser and Tcheng, 2006] [24], (b) utiliser l'indexation automatique des données non structurées [Hripcsak et al., 2009] afin de générer, dans le modèle, les EIs à partir des différents textes. Cette indexation est une solution pour structurer l'information afin de faciliter la RI.
Les approches sont nombreuses : [Rogers et al., 2006] implémentent une approche heuristique pour différencier, après extraction des concepts, une pathologie ponctuelle d'une pathologie récurrente, l'approche formelle de [Rector et al., 1993] pour différencier les observations des opinions faites sur celles-ci, la décomposition d'une phrase en blocs comme celle décrite par [Dogan et al., 2011], pour en déduire les relations sémantiques entre les concepts extraits ou la prise en compte des différentes rubriques [blocs] d'un document [Pinon et al., 1997; Pereira et al., 2009]. Cette dernière approche nous a permis de définir des métadonnées types "BLOC_CR" [25] dans notre modèle EI@DM afin de catégoriser les différents concepts extraits selon ses blocs et, donc de différencier une pathologie d'admission d'une pathologie antécédent. Le modèle EI@DM a été conçu pour pouvoir aussi gérer les informations provenant de l'indexation multi-terminologique des courriers et des comptes-rendus médicaux (métadonnées, valeurs numériques et symboliques, ...), qui seront de même format que les données structurées (codage PMSI, données biologiques).
Nous avons cité les outils TAL disponibles dans notre SI CISMeF tels que les outils `F-MTI` et `MCSVS`. L'évaluation de l'outil F-MTI [Pereira et al., 2009] pour l'indexation des documents médicaux non structurés a mis en évidence des insuffisances qui ne permettent pas de répondre à certaines requêtes sur les médicaments, aussi ne l'a-t-on pas implémenté lors de l'évalution. Les travaux dans le domaine sont moins abondants pour les documents en langue française. Les travaux de l'équipe CISMeF, sur l'indexation automatique des documents Web sont multi disciplinaire et multi lingue [Darmoni et al., 2009]. Une expérimentation des outils `F-MTI` [26] dans sa version améliorée [Pereira et al., 2011] et de `MCVS` sera réalisée dans le cadre du projet **RAVEL**.

24. L'ANR TecSan 2011 finance le projet **SIFaDo** (Saisie Informatique Facile de Données médicales) auquel l'équipe CISMeF participe. Ce projet vise à développer des méthodes et et des outils conduisant à des modalités de saisie de données utiles et faciles.

25. Nous avons défini un EI correspondant à un bloc d'un compte-rendu et qui sera indexé par l'ensemble des concepts extraits au niveau de ce bloc.

26. F-MTI est devenu un produit de la société VIDAL depuis

6.3.3 Limites du modèle

6.3.3.1 Temporalité de l'information

La prise en compte du contexte d'une observation médicale [27] impose une représentation de la temporalité [formelle ou non] afin de situer cet événement dans l'histoire médicale du patient. La problématique principale rencontrée durant notre évaluation est la non disponibilité de certaines informations liées à la temporalité des données structurées du DPI. Nous avons seulement pris en compte la temporalité par l'intermédiaire des dates sans développer des techniques de recherche et de fonctions spécifiques pour extraire ces informations dans les courriers et comptes-rendus médicaux.
Des choix de modélisation, comme l'ont décrit [Combi and Shahar, 1997], doivent être pris selon le besoin, afin de distinguer une date ponctuelle d'une durée ou d'avoir des intervalles de dates pour représenter l'incertitude des données cliniques [28] et traiter des données indépendantes du temps.
En effet, l'approche SPAN/SNAP du modèle CLEF [Rogers et al., 2006] est différente du concept HISTORY_EVENT définit par [Beale et al., 2007] pour modéliser la temporalité des informations cliniques. Le modèle openEHR distingue deux types de concepts : un concept INTERVAL_EVENT (une mesure réalisée dans un intervalle de temps comme le calcul d'une pression artérielle dans un intervalle de 5min) et un concept POINT_EVENT (un événement ponctuel comme le calcul du score d'APGAR). Un concept HISTORY_EVENT est un ensemble de POINT_EVENT et/ou d'INTERVAL_EVENT dans le modèle openEHR.
Sous réserve du développement au niveau des parseurs et des outils TAL des fonctions adaptées, il n'y a pas d'impossibilité à notre modèle à prendre en compte cette dimension. Ces fonctions vont prendre en compte les travaux déjà réalisés dans le domaine (l'approche SPAN/SNAP de [Rogers et al., 2006] ou les archétypes d'openEHR [Beale et al., 2007], ...).

6.3.3.2 Couverture du modèle

L'évaluation du modèle montre des limites pour certaines données (prescriptions médicamenteuses, poids, extubation, scores, paramètres viteaux [29], ...) afin de répondre à certaines questions des utilisateurs (l'évolution du poids d'un patient, les différentes valeurs des pressions systoliques et diastoliques, ...). Nous n'avons pas utilisé pour l'instant de modèle d'information mais pourrions dans le futur implémenter l'approche "Detailed Clinical Model (DCM)" en adoptant la modélisation à deux niveaux du méta modèle

27. Sens large du terme OSBERVATION (événement, état, fait, donnée clinique, ...)
28. Des données temporelles incomplètes
29. Un ensemble d'examens cliniques tels que la température du corps, ...

CISMeF. La particularité du DCM est d'être indépendant d'un modèle de références pour représenter un concept clinique. La revue de la littérature de [Gossens et al., 2010] résume cette approche DCM et compare différentes implémentations du DCM *(comme les templates HL7 RIM et les archétypes openEHR qui sont dépendants de leur modèle de référence).*

6.3.4 Intégration des données

HL7 est une représentation standard de communication des éléments de DPI. Cependant, la plupart des SIC ne l'implémente que partiellement. Face à l'émergence des technologies du Web Sémantique, plusieurs solutions existent pour exploiter les bases de données relationnelles (BDR) au travers des vues RDF [30]. Ces solutions utilisent l'approche ontologique pour le mapping entre les vues RDF et les BDR. Nous prévoyons de développer des parseurs OWL pour l'intégration des données dans une approche ontologique comme [Diallo, 2006; Mate et al., 2011].

30. W3C RDB2RDF Incubator Group [article] *"A Survey of Current Approaches for Mapping of Relational Databases to RDF"*, URL : http ://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport_01082009.pdf

Synthèse

Nous avons réalisé une validation applicative de notre modèle de données générique, concernant son aptitude à gérer les données du DPI et son adaptation à la RI tant au sein d'un seul dossier patient que dans une base multi-patients. Cette validation s'est faite à travers 2 éléments : l'utilisation des technologies sémantiques d'Oracle et le développement d'un prototype d'interfaces de RI **RIDoPI**. L'évaluation a révélé que la complexité des requêtes est identique tant pour les requêtes SQL simples que les requêtes SQL étendus avec des graphes RDF (SPARQL). Nous avons prouvé que des stratégies de RI (non actuellement implémentées) seraient nécessaires pour retrouver les informations médicales facilement, pouvoir gérer la temporalité de celles-ci et fournir ces informations sous une forme adaptée (résumé DPI, etc.). Nous avons observé, en plus, que notre prototype **RIDoPI** est *compliant* avec le framework **I2B2** d'un point de vue RI. Pour cela, des nouvelles métadonneés issues des données textuelles et d'autres données issues du SIC, dont l'intégration est simple dans notre modèle EI@DM, doivent être prises en compte pour répondre à plus de requêtes.

PERSPECTIVES

Dans ce chapitre, nous décrivons nos perspectives et nos projets de recherche en continuation avec le travail réalisé pendant cette thèse. Pour notre approche de modélisation et nos méthodes de RI, des améliorations sont nécessaires pour remédier aux limites identifiées et enrichir les approches adoptées ainsi que les implémentations mises en oeuvre.

Amélioration des travaux de thèse

En effet, l'évaluation du modèle EI@DM à travers les deux scénarios d'implémentation nous donne des perspectives prometteuses pour améliorer notre modélisation. Les améliorations à moyen terme vont concerner :

- la prise en compte de la chronologie pour raisonner sur certaines données temporelles et pouvoir faire des recherches événementielles plus complexes ;
- l'exploitation du contenu textuel du DPI en intégrant la dimension *chronologie* dans l'outil F-MTI [Pereira et al., 2009] qui prend en compte déjà les dimensions suivantes : le contexte, la prise en compte de la négation, l'extraction de concepts (les médicaments), ... ;
- l'enrichissement du dictionnaire de données de notre modèle ;
- l'amélioration des algorithmes de RI dans le prototype **RIDoPI**.

Pistes de réflexion et application

Cette thèse a clairement permis à l'équipe CISMeF de participer au consortium RAVEL en mai 2011, financé par l'ANR Programme TecSan .
L'évaluation des modèles d'information [de données] spécifiques au DPI déjà réalisée, en partie dans cette thèse, sera valorisée avec une étude plus approfondie dans le cadre de ce projet afin de déterminer le dénominateur commun entre ces différents modèles et leur utilisation pour la recheche d'information et la visualisation.
Un passage à l'échelle sera réalisé à partir de la *preuve de concept* réalisée dans cette

thèse avec les technologies sémantiques d'Oracle, afin d'évaluer la plus-value du Web Sémantique (la scalabilité du SPARQL, la formalisation OWL et les outils d'inférences) avec les outils CISMeF. Nous évaluerons le corpus test de [Koopman et al., 2011] sur la version améliorée du protoype **RIDoPI**.
Nous avons vu, dans notre travail, l'importance d'au moins 4 terminologies (CIM10, CCAM, LOINC, ATC) pour la RI dans un DPI. Nous testerons l'interêt d'interfacer d'autres terminologies, comme celles sur les dispositifs médicaux CISP, Cladimed. Surtout nous travaillerons sur les terminologies d'interfaces (en particulier de prescription) pour améliorer la RI dans le DPI.
A plus long terme, une RI dans un dossier personnel tendant vers une médecine personnalisée alliant phénotype et génotype, nous semble une visée de recherche prometteuse pour l'équipe TIBS qui associent la BioInformatique et l'Informatique Médicale. Cette médecine personnalisée représente l'axe principal de recherche de l'équipe TIBS dans les dix prochaines années.

RIDoPI2@RAVEL

Cette thèse est un tournant stratégique pour l'équipe CISMeF qui s'est focalisée sur la RI dans un catalogue de ressources et non au sein d'un DPI. Nous avons vu dans ce travail que la modélisation de cette RI avait nécessité de prendre en compte des données numériques (en particulier la biologie) et des données chronologiques. Le sujet de cette thèse va se poursuivre dans le cadre du projet RAVEL. Ce projet démarre en Janvier 2012 durant 3 ans. Il va permettre de réutiliser le moteur de recherche Doc'CISMeF dans ce nouveau contexte, en s'appuyant sur les travaux sur la modélisation menée dans cette thèse. Cette thèse sera suivi par deux autres sujets de thèse trés connexes à celle-ci qui ont toutes les deux débutées en Octobre 2011 :

1. une thèse sur les terminologies d'interface (Nicolas GRIFFON) en particulier en biologie et en radiologie. Ces travaux permettront d'améliorer grandement la RI dans le DPI car la plupart des médecins ne connaissent de façon approfondie ni LOINC, ni le CCAM dans un contexte de RI. Les terminologies d'interface s'adapteront au langage des professionels de santé, même s'il est ambigu notamment du fait de l'utilisation d'acronymes ;
2. la modélisation d'un outil de gestion et de navigation cross-lingue autour des terminologies et ontologies de santé (Julien GROSJEAN). [Lovis et al., 2011] préconisent le développement d'outils multilingues d'indexation et d'extraction des données du DPI pour leur re-utilisation mais aussi pour rendre interopérable plusieurs DPI de pays et de cultures différents.

CONCLUSION GENERALE

Notre problématique principale était de concevoir un modèle sémantique de données pour représenter l'hétérogénéité et la complexité des données médicales du DPI, afin d'avoir une information sous une forme *exploitable* par un système de recherche d'information.
Pour ce faire, nous avons développé le modèle EI@DM : un modèle générique et flexible pour ne pas se limiter aux données médicales issues de la base de données CDP du CHU de Rouen et, un modèle sémantiquement riche à travers les terminologies et les ontologies se trouvant dans le PTS.
Au sein du SI CISMeF, nous avons développé un prototype d'interfaces de RI : **RIDoPI**, un outil de RI mono et multi-patients. Les résultats de l'étude que nous avons menée pour évaluer la conformité fonctionnelle de notre prototype **RIDoPI** avec le *framework* **I2B2**, nous ont révélé que notre prototype est *compliant* avec ce framework d'un point de vue RI. L'échantillon utilisé n'est évidemment pas représentatif des requêtes qui peuvent se poser dans un dossier patient mais nous avons pu formuler, dans 90% des cas, les requêtes événementielles à travers nos interfaces de RI et ce qui est prometteur pour l'évolution de ce prototype.
Notre objectif dans le futur est d'appliquer les pistes d'amélioration pour répondre à plus de requêtes sans modification de notre modèle de données, ni de nos méthodes de RI, afin de répondre aux besoins des utilisateurs dans des délais brefs.

Liste des Publications

1. Griffon, N., Sakji, S., Dirieh Dibad, A.D, Grosjean, J., Philippe, M. et Darmoni, S.J. A model for Information Retrieval in Electronic Health Records. In *Proceedings of the 3rd International Workshop on Knowledge Representation for Health Care (KR4HC'11) in conjunction with the 13th Conference on Artificial Intelligence in Medecine (AIME'11)*, pages 165–170.
2. Dirieh Dibad, A.D, Griffon, N., Sakji, S., Pereira, S., Massari, P., and Darmoni, S.J. Information retrieval in electronic health record : a feasibility study. In *Proceedings of the 23th International Conference of the European Federation for Medical Informatics (MIE'11)* (posters).
3. Dirieh Dibad, A.D, Soualmia, L.F, Merabti, T., Grosjean, J., Sakji, S., Massari, P., and Darmoni, S.J. Un modèle de données adapté à la recherche d'information dans le dossier patient informatisé : étude, conception et évaluation. In *Systèmes d'information pour l'amélioration de la qualité en santé : Comptes-Rendus des 14^{e} Journées Francophones d'Informatique Médicale (JFIM'11)*, volume 18 of *Informatique et Santé*, pages 251–262.
4. Sakji, S., Dirieh Dibad, A.D, Kergourlay, I., Joubert, M. et Darmoni, S.J. Information Retrieval in Context Using Various Health Terminologies. In *Proceedings of the 3rd International Conference on Research Challenges in Information Science (RCIS'09)*, pages 453–458, IEEE. PSIP
5. Dirieh Dibad, A.D, Sakji, S., Prieur, E., Pereira, S., Joubert, M., and Darmoni, S.J. Recherche d'information multi-terminologique en contexte : Etude préliminaire. In *Risques, technologies de l'information pour les pratiques médicales : Comptes-Rendus des 13^{e} Journées Francophones d'Informatique Médicale (JFIM'09)*, volume 17 of *Informatique et Santé*, pages 101–112. PSIP.

Bibliographie

Ammenwerth, E., Hackl, W., Massari, P., and Darmoni, S. (2011). Validation of completeness, correctness, relevance and understandability of the psip cdss for medication safety. In *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, volume 166 of *Studies in Health Technology and Informatics*, pages 254–259.

ANAP (2010). Modèles conceptuels de santé. Rapport technique, Appui Santé & médico-social.

Anderson, J. (2007). Social, ethical and legal barriers to e-health. *International Journal of Medical Informatics*, 76(5-6) :480–483.

Arguello, M., Des, J., Perez, R., et al. (2009). Electronic health records (ehrs) standards and the semantic edge : A case study of visualising clinical information from ehrs. In *Proceedings of the 11th International Conference on Computer Modelling and Simulation (UKSiM'09)*, pages 485–490. IEEE.

Austin, T., Kalra, D., Tapuria, A., et al. (2008). Implementation of a query interface for a generic record server. *International Journal of Medical Informatics*, 77(11) :754–764.

Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999). *Modern information retrieval*, volume 82. Addison-Wesley.

Barnett, G. (2006). Report to the national institutes of health division of research grants computer research study section on computer applications in medical communication and information retrieval systems as related to the improvement of patient care and the medical record. *Journal of the American Medical Informatics Association*, 13(2) :127–135.

Barrows Jr, R. and Johnson, S. (1995). A data model that captures clinical reasoning about patient problems. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 402. AMIA.

Baud, R., Lovis, C., and Scherrer, J. (1998). *Nouvelles perspectives en matière de dossiers patients en réseau*, volume 10 of *Informatique et Santé*, pages 45–55. Springer-Verlag.

Bayegan, E., Nytro, O., and Grimsmo, A. (2002). Ontologies for knowledge representation in a computer-based patient record. In *Proceedings of the 14th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 114–121. IEEE.

Beale, E. T., Heard, S., Kalra, D., and Lloyd, D. (2007). The openehr reference model - ehr information model. *Health San Francisco*.

Beale, T. (2002). Archetypes : Constraint-based domain models for futureproof information systems. In *Proceeding of the 11th The International Conference on Object Oriented Programming, Systems, Languages and Applications (OOPSLA) Workshop on Behavioral Semantics*, pages 1–18.

Beneventano, D., Bergamaschi, S., Castano, S., et al. (2000). Information integration : the momis project demonstration. In *Proceedings of the International Conference on Very Large Data Bases (VLDB'00)*, pages 611–614.

Bird, L., Goodchild, A., and Tun, Z. (2003). Experiences with a two-level modelling approach to electronic health records. *Journal of Research and Practice in Information Technology*, 35(2) :121–138.

Bodenreider, O., Smith, B., Kumar, A., and Burgun, A. (2007). Investigating subsumption in snomed ct : An exploration into large description logic-based biomedical terminologies. *Artificial intelligence in medicine*, 39(3) :183–195.

Boonstra, A. and Manda, B. (2010). Barriers to the acceptance of electronic medical records by physicians from systematic review to taxonomy and interventions. *BMC Health Services Research*, 10(1) :231.

Brachman, R. and Schmolze, J. (1985). An overview of the kl-one knowledge representation system. *Cognitive science*, 9(2) :171–216.

Branson, A., Hauer, T., McClatchey, R., et al. (2008). A data model for integrating heterogeneous medical data in the health-e-child project. *Studies in Health Technology and Informatics*, 138 :13–23.

Braun, L., Wiesman, F., van den Herik, H., et al. (2007). Towards patient-related information needs. *International Journal of Medical Informatics*, 76(2) :246–251.

Braun, L. M. M. (2008). *Pro-Active Medical Information Retrieval*. Thèse de doctorat, Université de Maastricht.

Bréchot, C. (2004). La recherche translationnelle en santé, un nouveau paradigme. *Médecine/Sciences*, 20(10) :939–940.

Callebat, L. (1999). Le médecin au moyen-âge. In *Histoire du médecin*, pages 59–109. Flammarion.

Campbell, K., Das, A., and Musen, M. (1994). A logical foundation for representation of clinical data. *Journal of the American Medical Informatics Association*, 3(1) :218–232.

Cao, Y., Cimino, J., Ely, J., and Yu, H. (2010). Automatically extracting information needs from complex clinical questions. *Journal of Biomedical Informatics*, 43(6) :962–971.

CAP (2006). Snomed clinical terms guide. Rapport technique, College of American Pathologists.

Choquet, R., Daniel, C., Boussaid, O., and Jaulent, M. (2008). Etude méthodologique comparative de solutions d'entreposage de données de santé à des fins décisionnelles. In *Proceedings of the International Conference on System Science in Health Care (ICSSHC'08)*, pages 1–6.

Christensen, T. and Grimsmo, A. (2008). Instant availability of patient records, but diminished availability of patient information : A multi-method study of gp's use of electronic patient records. *BMC medical informatics and decision making*, 8(1) :12.

Chute, C. and Yang, Y. (1992). An evaluation of concept based latent semantic indexing for clinical information retrieval. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 639. AMIA.

Cimino, J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, 37(4-5) :394–403.

Cimino, J. (2006). In defense of the desiderata. *Journal of Biomedical Informatics*, 39(3) :299–306.

Cimino, J. J. and Jianhua, L. (2003). Sharing infobuttons to resolve clinicians'information needs. In *AMIA Annu Symp Proc.*, volume 2003, page 815.

Cimino, J. J., Johnson, S. B., Aguirre, A., Roderer, N., and Clayton, P. D. (1992). The MEDLINE button. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 81–85. AMIA.

CNEH (2011). Guide relatif aux modalités d'accès au dossier médical du patient. Rapport technique, Centre national de l'expertise hospitalière (CNEH).

Collen, M. (1991). A brief historical overview of hospital information system (his) evolution in the united states. *International journal of bio-medical computing*, 29(3-4) :169–189.

Collen, M. F. (1990). Clinical research databases–a historical review. *Journal of Medical Systems*, 14(6) :323–344.

Combi, C. and Shahar, Y. (1997). Temporal reasoning and temporal data maintenance in medicine : issues and challenges. *Computers in Biology and Medicine*, 27(5) :353–368.

Cuggia, M., Bayat, S., Garcelon, N., Sanders, L., Rouget, F., Coursin, A., and Pladys, P. (2010). A full-text information retrieval system for an epidemiological registry. *Studies In Health Technology And Informatics*, 160(Pt 1) :491–495.

Cuggia, M., Bayat, S., Rossille, D., et al. (2009). Comparing the apgar score representation in hl7 and openehr formalisms. *Studies in Health Technology and Informatics*, 150 :250–254.

Cuggia, M., Garcelon, N., Campillo-Gimenez, B., Bernicot, T., Laurent, J., Garin, E., Happe, A., and Duvauferrier, R. (2011). Roogle : an information retrieval engine for clinical data warehouse. *Studies in Health Technology and Informatics*, 169 :584–588.

Currie, A., Cohan, J., and Zlatic, L. (2001). Information retrieval of electronic medical records. *Computational Linguistics and Intelligent Text Processing*, pages 460–471.

Daniel, C., Jais, J., Fadly, N., and Landais, P. (2009). Dossier patient informatisé à visée de recherche biomédicale. *La Presse Médicale*.

Darmoni, S., Sakji, S., Pereira, S., T, M., E, P., M, J., and B, T. (2009). Multiple terminologies in an health portal : automatic indexing and information retrieval. In *Artificial Intelligence in Medicine*, Lecture Notes in Computer Science, pages 255–259. Springer. PSIP. InterSTIS.

Darmoni, S., Thirion, B., Leroy, J., and Douyère, M. (2001a). The use of dublin core metadata in a structured health resource guide on the internet. *Bulletin Medical Library Association*, 89(3) :297–301.

Darmoni, S., Thirion, B., Leroy, J. P., Douyère, M., Lacoste, B., Godard, C., Rigolle, I., Brisou, M., Videau, S., Goupy, E., Piot, J., Quéré, M., Ouazir, S., and Abdulrab, H. (2001b). Doc'cismef : a search tool based on "encapsulated" mesh thesaurus. *Studies in health technology and informatics*, 10(Pt 1) :314–318.

Degoulet, P. (1984). *L'informatisation du dossier médical : les axes sémantiques et temporels*. Thèse de doctorat, Université Pierre et Marie-Curie.

Degoulet, P. and Fagon, J. (2004). Stratégies de mise en oeuvre des systèmes d'information cliniques. *Gestions Hospitalières*, 2004(441) :793–800.

Degoulet, P. and Fieschi, M. (1991). *Traitement de l'information médicale : méthodes et applications hospitalières*. Masson.

Degoulet, P. and Jean, F. (1989). The need for pragmatic data models. In *Proceeding of Computerized Natural Language Processing for Knowledge Engineering*, pages 157–167.

Degoulet, P., Marin, L., Kleinebreil, L., and Albigès, L. (2003). *Présent et avenir des systèmes d'information et de communication hospitalier (SICH)*, volume 15 of *Informatique et Santé*. Springer-Verlag.

DeLisle, S., South, B., Anthony, J. A., et al. (2010). Correction : Combining free text and structured electronic medical record entries to detect acute respiratory infections. *Public Library of Science (PLoS) ONE*, 5(10) :9.

Deshmukh, V., Meystre, S., and Mitchell, J. (2009). Evaluating the i2b2 system for clinical research. *BMC Medical Research Methodology*, 9(1) :70.

Diallo, G. (2006). *Une architecture à base d'ontologies pour la gestion unifiée des données structurées et non structurées*. Thèse de doctorat, Université Joseph Fourrier (Grenoble I).

Dieng-Kuntz, R., Corby, O., Gandon, F., Golebiowska, G., Matta, N., and Ribière, M. (2001). *Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du KnowledgeManagement*. 2ème Edition. Dunod.

Dirieh Dibad, A., Griffon, N., Sakji, S., Pereira, S., Massari, P., and Darmoni, S. (2011a). Information retrieval in electronic health record : a feasibility study. In *Proceedings of the 23th International Conference of the European Federation for Medical Informatics (MIE'11)*.

Dirieh Dibad, A., Sakji, S., Prieur, E., Pereira, S., Joubert, M., and Darmoni, S. (2009). Recherche d'information multi-terminologique en contexte : Etude préliminaire. In *Risques, technologies de l'information pour les pratiques médicales : Comptes-Rendus des 13e Journées Francophones d'Informatique Médicale (JFIM)*, volume 17 of *Informatique et santé*, pages 101–112. Springer. PSIP.

Dirieh Dibad, A., Soualmia, L., Merabti, T., Grosjean, J., Sakji, S., Massari, P., and Darmoni, S. (2011b). Un modèle de données adapté à la recherche d'information dans le dossier patient informatisé : étude, conception et évaluation. In *Systèmes*

d'information pour l'amélioration de la qualité en santé : Comptes-Rendus des 14^{e} Journées francophones d'informatique médicale (JFIM'11), volume 18 of *Informatique et Santé*, pages 251–262.

Dogan, R. I., Neveol, A., and Lu, Z. (2011). A context-blocks model for identifying clinical relationships in patient records. *BMC Bioinformatics*, 12(Suppl 3) :S3.

Dolin, R. (1994). A high-level object-oriented model for representing relationships in an electronic medical record. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 514. AMIA.

Dolin, R. (1995). Modeling the temporal complexities of symptoms. *Journal of the American Medical Informatics Association*, 2(5) :323.

Dolin, R., Alschuler, L., Beebe, C., Biron, P., Boyer, S., Essin, D., Kimber, E., Lincoln, T., and Mattison, J. (2001). The hl7 clinical document architecture. *Journal of the American Medical Informatics Association*, 8(6) :552.

Dolin, R., Alschuler, L., Boyer, S., Beebe, C., Behlen, F., Biron, P., and Shabo Shvo, A. (2006). Hl7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1) :30–39.

Douyère, M., Soualmia, L., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J., Thirion, B., and Darmoni, S. (2004). Enhancing the mesh thesaurus to retrieve french online health resources in a quality-controlled gateway. *Health Information and Libraries Journal*, 21(4) :253–261.

Eichelberg, M., Aden, T., Riesmeier, J., Dogac, A., and Laleci, G. (2005). A survey and analysis of electronic healthcare record standards. *ACM Computing Surveys*, 37(4) :277–315.

Embi, P. J., Jain, A., and Harris, C. M. (2008). Physicians' perceptions of an electronic health record-based clinical trial alert approach to subject recruitment : A survey. *BMC Medical Informatics and Decision Making*, 8(1) :13.

Essin, D. and Lincoln, T. (1994). An information model for medical events. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 509. AMIA.

Fakoff, H. (1999). *Le dossier orienté problème existe, je l'ai rencontré*, volume 11 of *Informatique et Santé*, chapter Informatisation du Cabinet Médical du Futur, pages 149–157. Springer-Verlag.

Farfan, F., Hristidis, V., Ranganathan, A., and Weiner, M. (2009). Xontorank : Ontology-aware search of electronic medical records. In *Proceedings of the 25th International Conference on Data Engineering*, pages 820–831. IEEE.

Feldman, R. and Goodrich, J. (1999). The edwin smith surgical papyrus. *Child's Nervous System*, 15(6) :281–284.

Fieschi, M. (2003). Les données du patient partagées : la culture du partage et de la qualité des informations pour améliorer la qualité des soins. Rapport technique, ASIP Santé.

Flexner, A. (1910). *Medical Education in the United States and Canada.* Carnegie Foundation for the Advancement of Teaching.

Flory, A., Laforest, F., and Weill, A. (2000). Utilisation des documents semi-structurés pour la représentation et le stockage du dossier médical. *L'informatique au service du patient*, 12 :229–240.

Flory, A., Verdier, C., and Sassi, S. (2006). Nouvelles interfaces pour la représentation de l'information médicale. In *INFORSID*, pages 177–197.

Fraser, H., Biodich, P., Moodley, D., Choi, S., Mamlin, B., and Szolovits, P. (2005). Implementing electronic medical records systems in developing countries. *Informatics in Primary Care*, 14(2).

Ganslandt, T., S., M., Helbing, K., Sax, U., and Prokosch, H. (2011). Unlocking data for clinical research - the german i2b2 experience. *Applied Clinical Informatics*, 2.

Gardner, M. (1997). Information retrieval for patient care. *British Medical Journal (BMJ)*, 314(7085) :950.

Gehanno, J., Thirion, B., and Darmoni, S. (2007). Evaluation of meta-concepts for information retrieval in a quality-controlled health gateway. In *AMIA Symp.*, pages 269–73. IOS Press.

Goble, C., Bechhofer, S., Solomon, W., Rector, A., Nowlan, W., and Glowinski, A. (1994). Conceptual, semantic and information models for medicine. In *Proceedings of the 4th European-Japanese Seminar on Information Modelling and Knowledge Bases*, pages 257–286.

González, A., Dawes, M., Sánchez-Mateos, J., Riesgo-Fuertes, R., Escortell-Mayor, E., Sanz-Cuesta, T., and Hernández-Fernández, T. (2007). Information needs and information-seeking behavior of primary care physicians. *The Annals of Family Medicine*, 5(4) :345–352.

Gossens, W., Gossens-Baremans, A., and Van Der Zeln, M. (2010). Detailed clinical models : a review. *Healthcare Informatics Researach (HIR)*, 16(4) :201–214.

Gouveia-Oliveira, A. and Lopes, L. (1993). Formal representation of a conceptual data model for the patient-based medical record. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 466. AMIA.

Grenon, P. and Smith, B. (2004). SNAP and SPAN : Towards Dynamic Spatial Ontology. *Spatial Cognition & Computation : An Interdisciplinary Journal*, 4(1) :69–104.

Griffon, N., Sakji, S., Dirieh Dibad, A., Grosjean, J., Philippe, M., and Darmoni, S. (2011). A model for information retrieval in electronic health records. In *Proceedings of the 3rd International Workshop on Knowledge Representation for Health Care (KR4HC'11) in conjunction with the 13th Conference on Artificial Intelligence in Medecine (AIME'11)*, pages 165–170.

Grosjean, J., Merabti, T., Dahamna, B., Kergourlay, I., B, T., LF, S., and Darmoni, S. (2011a). Health multi-terminology portal : a semantics added-value for patient safety. In *Patient Safety Informatics - Adverse Drug Events, Human Factors and IT Tools for Patient Medication Safety*, volume 166 of *Studies in Health Technology and Informatics*, pages 129–138. PSIP.

Grosjean, J., Merabti, T., Griffon, N., Dahamna, B., and SJ, D. (2011b). Multiterminology cross-lingual model to create the european health terminology/ontology portal. In *Proceedings the 9th International Conference on Terminology and Artificial Intelligence(TIA)*, pages 119–122.

Guisiano, B., Ledoray, V., Jimeno, M., and Flory, A. (1992). *Sémantique du dossier médical et bases de données*, volume 5 of *Informatique et Santé*, pages 18–32. Springer-Verlag.

Guo, L., Shao, F., Botev, C., and Shanmugasundaram, J. (2003). Xrank : Ranked keyword search over xml documents. In *Proceedings of the International Conference on Management of data (ACM/SIGMOD'03)*, pages 16–27.

Hansen, M., Madnick, S., and Siegel, M. (2003). Data integration using web services. *Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web*, pages 165–182.

Hasman, A., Safran, C., and Takeda, H. (2003). Quality of health care : Informatics foundations. *Methods of Information in Medicine*, 42(5) :509–518.

Haux, R. (2006). Health information systems - past, present, future. *International Journal of Medical Informatics*, 75(3-4).

Haux, R., Ammenwerth, E., Herzog, W., and Knaup, P. (2002). Health care in the information society. a prognosis for the year 2013. *International Journal of Medical Informatics*, 66(1–3) :3–21.

Hersh, W. (2008). *Information Retrieval : A Health and Biomedical Perspective*. Health Informatics. Springer, 3ıème edition.

Hersh, W. and Hickam, D. (1998). How well do physicians use electronic information retrieval systems ? *JAMA : the journal of the American Medical Association*, 280(15) :1347.

HL7 (2005). Hl7 reference information model (rim). Rapport technique, Health Level Seven (HL7).

HPRIM, H. (2009). Guide d'implémentation de l'entête de documents cda (version 1.0). Rapport technique, Interop'Santé.

Hripcsak, G., Soulakis, N. D., Li, L., Morrison, F. P., Lai, A. M., Friedman, C., Calman, N. S., and Mostashari, F. (2009). Syndromic surveillance using ambulatory electronic health records. *Journal of the American Medical Informatics Association*, 16(3) :354–361.

Hripcsak, G., Zhou, L., Parsons, Sand Das, A., and Johnson, S. (2005). Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. *Journal of the American Medical Informatics Association*, 12(1) :55–63.

Hristidis, V., Varadarajan, R., Biondich, P., and Weiner, M. (2010). Information discovery on electronic health records using authority flow techniques. *BMC Medical Informatics and Decision Making*, 10(1) :64.

Huet, B., Lesueur, B., Artigou, J., and Blaint, G. (2001). Méta-modélisation du dossier médical conception, intérêt, application. *JFIM*, 12 :217–228.

Huff, S., Rocha, R., Bray, B., Warner, H., and Haug, P. (1995). An event model of medical information representation. *Journal of the American Medical Informatics Association*, 2(2) :116.

Hull, R. and King, R. (1987). Semantic database modeling : survey, applications, and research issues. *ACM Computing Surveys (CSUR)*, 19(3) :201–260.

Häyrinen, K., Saranto, K., and Nykänen, P. (2008). Definition, structure, content, use and impacts of electronic health records : a review of the research literature. *International Journal of Medical Informatics*, 77(5).

Iakovidis, I. (1998). Towards personal health record : current situation, obstacles and trends in implementation of electronic healthcare record in europe. *International Journal of Medical Informatics*, 52(1-2) :105–115.

ISO, . (1998). Iso 14258 : Industrial automation systems - concepts and rules for enterprise models. Rapport technique, ISO.

ISO, T. . (2005). Health informatics - electronic health record - definition , scope , and context. Rapport technique, ISO.

Jain, H., Thao, C., and Zhao, H. (2010). Enhancing electronic medical record retrieval through semantic query expansion. *Information Systems and E-Business Management*, pages 1–17.

Johnson, S. (1996). Generic data modeling for clinical rrepositories. *Journal of the American Medical Informatics Association*, 3(5) :328.

Johnson, S., Friedman, C., Cimino, J., Clark, T., Hripcsak, G., and Clayton, P. (1991). Conceptual data model for a central patient database. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 381. AMIA.

Joubert, M., Robert, J., Miton, F., and Fieschi, M. (1996). The project ariane : conceptual queries to information databases. In *Proceedings of the AMIA Annual Fall Symposium*, page 378. AMIA.

Kahn, M. (1991). Modeling time in medical decision-support programs. *Medical Decision Making*, 11(4) :249.

Kalra, D. and Lloyd, D. (1995). The gehr final architecture description. Rapport technique, European Commission.

Kamadjeu, R., Tapang, E., and Moluh, R. (2005). Designing and implementing an electronic health record system in primary care practice in sub-saharan africa : a case study from cameroon. *Informatics in primary care*, 13(3) :179–186.

Kanter, A., Wang, A., Masarie, F., Naeymi-Rad, F., and Safran, C. (2008). Interface terminologies : bridging the gap between theory and reality for africa. *Studies in health technology and informatics*, 136 :27.

Kay, S. and Marley, T. (1999). *ENV 13606, ECHR Communications, Part 1 Electronic Healthcare Record Architecture*. CEN TC/251.

Ken, S. and Dipak, K. (2008). Electronic health records in complementary and alternative medicine. *International Journal of Medical Informatics*, 77 :576–588.

Kohl, C., Garde, S., and Knaup, P. (2010). Facilitating secondary use of medical data by using openehr archetypes. *Studies in Health Technology Informatics*, 160(Pt 2) :1117–21.

Koopman, B., Bruza, P., Sitbon, L., and Lawley, M. (2011). *Evaluating Medical Information Retrieval*, pages 1139–1140. ACM.

Lai, A., Parsons, S., and Hripcsak, G. (2008). Fuzzy temporal constraint networks for clinical information. In *Proceedings of the AMIA Annual Symposium*, volume 2008, page 374. AMIA.

Lamy, J., Duclos, C., Hamek, S., et al. (2010). Towards iconic language for patient records, drug monographs, guidelines and medical search engines. *Studies in Health Technology and Informatics*, 160(pt 1) :156–160.

Lehmann, T. and Meyer zu Bexten, E. (2002). *Handbuch der medizinischen Informatik*. Hanser Verlag.

Lenz, R., Beyer, M., and Kuhn, K. (2007). Semantic integration in healthcare networks. *International journal of medical informatics*, 76(2-3) :201–207.

Lieberman, D. (2010). Pitfalls of using administrative data for research. *Digestive diseases and sciences*, 55(6) :1506–1508.

Lindemann, G., Schmidt, D., Schrader, T., and Keune, D. (2009). The resource description framework (rdf) as a modern structure for medical data. *International Journal of Biological and Medical Sciences*, 4(2).

Liu, S., Ni, Y., Mei, J., Li, H., Xie, G., Hu, G., Liu, H., Hou, X., and Pan, Y. (2009). ismart : Ontology-based semantic query of cda documents. In *Proceedings of the AMIA Annual Symposium*, volume 2009, page 375. AMIA.

Los, R., Roukema, J., van Ginneken, A., de Wilde, M., and van der Lei, J. (2005). Are structured data structured identically ? investigating the uniformity of pediatric patient data recorded using opensde. *Methods of Information in Medicine*, 44(5).

Lovis, C., Ball, M., Boyer, C., and Elkin, P. (2011). Hospital and health information systems - current perspectives. contribution of the imia health information systems working group. *Yearbook of Medical Informatics*, 1 :73–82.

Lowe, H., Ferris, T., Hernandez, P., and Weber, S. (2009). Stride–an integrated standards-based translational research informatics platform. In *Proceedings of the AMIA Annual Symposium*, volume 2009, page 391. AMIA.

Lukacs, B. and Lang, A. (1989). *Les grandes fonctionnalités d'un système de gestion de l'unité de soins*, volume 1 of *Informatique et Santé*, pages 15–24. Springer-Verlag.

Lyman, J., Pelletier, S., Scully, K., et al. (2003). Applying the hl7 reference information model to a clinical data warehouse. In *Proceedings of IEEE International Conference on the Systems, Man and Cybernetics*, volume 5, pages 4249–4255. IEEE.

Ma, C., Frankel, H., Beale, T., and Heard, S. (2007). Ehr query language (eql)- a query language for archetype-based health records. *Studies In Health Technology And Informatics*, 129(Pt 1) :397–401.

Macary, F. (2007). Ihe, cda et loinc : des composants d'interoperabilite au service du partage des resultats de biologie medicale. *Spectra biologie*, 158 :51.

Marc, J., Michel, R., François, M., and Jacques, H. (2009). Origine de la cisp et mise en application actuelle dans les pays francophones. In *L'Informatisation du Cabinet Médical du Futur*, volume 11 of *Informatique et Santé*, pages 201–212.

Massari, P. and Fuss, J. (2000). Dossier patient informatisé du chu de rouen : migration des anciennes applications vers c_page dossier patient. *Gestions Hospitalères*, 2000(395) :316–320.

Massari, P., Pereira, S., Thirion, B., Derville, A., and Darmoni, S. (2008). Use of superconcepts to customize electronic medical records data display. In *Beyond the Horizon - Get IT There - Proceedings of MIE2008 - The XXIst International Congress of the European Federation for Medical Informatics*, volume 136 of *Studies in Health Technology and Informatics*, pages 845 – 850.

Mate, S., Burkle, T., Prokosh, H., and Ganslandt, T. (2011). Populating the i2b2 database with data from the electronic medical record : Ontology-based form data integration for i2b2. In *MIE2011*.

McDonald, C., Huff, S., Suico, J., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., et al. (2003). Loinc, a universal standard for identifying laboratory observations : A 5-year update. *Clinical Chemistry*, 49 :624–633.

Merabti, T. (2010). *Méthodes pour la mise en relation des terminologies médicales : contribution à l'interopérabilité sémantique Inter et Intra terminologique.* Thèse de doctorat, Université de Rouen.

Merabti, T., Abdoune, H., Lecroq, T., Joubert, M., and Darmoni, S. (2009). Projection des relations snomed ct entre les termes de deux terminologies (cim10 et snomed 3.5). In *Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 of *Informatique et santé*, pages 79–88.

Merabti, T., Grosjean, J., Abdoune, H., and Joubert, M. a. D. S. (2011). Automatic methods for mapping biomedical terminologies in a health multi-terminology portal. In *EGC 2011 : Atelier Extraction de Connaissances et Santé.*

Moutel, G. (2009). *Médecins et Patients : L'exercice de la démocrtie sanitaire.* L'Ethique en mouvement. L'Harmattan.

Müller, H., Deselaers, T., Deserno, T., Kalpathy-Cramer, J., Kim, E., and Hersh, W. (2008). Overview of the imageclefmed 2007 medical retrieval and medical annotation tasks. *Advances in Multilingual and Multimodal Information Retrieval*, pages 472–491.

Murff, H., FitzHenry, F., Matheny, M., Gentry, N., Kotter, K., Crimin, K., Dittus, R., Rosen, A., Elkin, P., Brown, S., et al. (2011). Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA : The Journal of the American Medical Association*, 306(8) :848–855.

Murphy, S., Mendis, M., Berkowitz, D., Kohane, I., and Chueh, H. (2006). Integration of clinical and genetic data in the i2b2 architecture. In *Proceedinfs of the AMIA Annual Symposium*, volume 2006, page 1040. AMIA.

Musser, C. R. and Tcheng, J. E. (2006). Quantitative and qualitative comparison of text-based and graphical user interfaces for computerized provider order entry. In *Proceedings of the AMIA Annual Symposium*, volume 2006, page 1041. AMIA.

Nadkarn, P., Marenco, L., Chen, R., Skoufos, E., and Shepherd, G. (1999). Organization of heterogeneous scientific data using the eav/cr representation. *Journal of the American Medical Informatics Association*, 6(6) :478–493.

Natarajan, K., Stein, D., Jain, S., and Elhadad, N. (2010). An analysis of clinical queries in an electronic health record search utility. *International Journal of Medical Informatics*, 79(7) :515–522.

Niso, P. (2004). Understanding metadata. *National Information Standards.*

Nowlan, W. (1993). *Structured Methods of Information Management for Medical Records.* Thèse de doctorat, University of Manchester.

Névéol, A. (2005). *Automatisation des tâches documentaires dans un catalogue de santé en ligne.* Thèse de doctorat, INSA de Rouen.

Nygren, E. and Henriksson, P. (1992). Reading the medical record. i. analysis of physician's ways of reading the medical record. *Computer Methods and Programs in Biomedicine*, 39(1-2) :1–12.

Olesen, H. (1996). Properties and units in the clinical laboratory sciences. i. syntax and semantic rules iupac–ifcc recommendations 1995. *Clinica chimica acta international journal of clinical chemistry*, 245(2) :S5–S21.

Ondo, K., Wagner, J., and Gale, K. (2002). The electronic medical record : Hype or reality ? *Journal of Healthcare Information Management*, 17(4) :2.

Osheroff, J., Forsythe, D., Buchanan, B., Bankowitz, R., Blumenfeld, B., and Miller, R. (1991). Physicians' information needs : analysis of questions posed during clinical teaching. *Annals of Internal Medicine*, 114(7) :576–581.

Patel, C. and Cimino, J. (2007). Matching patient records to clinical trials using ontologies. In *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 816–829. SPRINGER.

Pereira, S. (2008). *Indexation automatique multi terminologique.* Thèse de doctorat, Université de Rouen.

Pereira, S., Letord, C., Darmoni, S., and Serrot, E. (2011). Extraction des noms de médicaments dans les comptes rendus hospitaliers. In *Systèmes d'information pour l'amélioration de la qualité en santé. Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM)*, Informatique et Santé, pages 145–154. Springer.

Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M., and S.J., D. (2009). F-mti : outil d'indexation multi-terminologique : application à l'indexation automatique de la snomed. In *Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 of *Informatique et santé*, pages 57–67. Springer-Verlag.

Pereira, S., Névéol, A., Massari, P., Joubert, M., and Darmoni, S. (2006). Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. *Studies in Health Technology and Informatics*, 124 :845–850.

Pinon, J., Calabretto, S., and Poullet, L. (1997). Document semantic model : An experiment with patient medical record. In *Proceedings of the conference on Electronic Publishing, Canterbury*, pages 262–272.

Plaza, L. and Díaz, A. (2010). Retrieval of similar electronic health records using umls concept graphs. *Natural Language Processing and Information Systems*, pages 296–303.

Pouliquen, B. (2002). *indexation de textes médicaux par extraction de concepts et ses utilisation*. Thèse de doctorat, Université de Rennes 1, Faculté de Médecine.

Powell, J. and Buchan, I. (2005). Electronic health records should support clinical research. *Journal of Medical Internet Research*, 7(1) :e4.

Powsner, S., Tufte, E., et al. (1994). Graphical summary of patient status. *Lancet*, 344(8919) :386–389.

Price, S. L., Hersh, W. R., Olson, D. D., and Embi, P. J. (2002). Smartquery : Context-sensitive links to medical knowledge sources from the electronic patient record. In *In Proceedings of AMIA Annual Symp*.

Qamar, R. and Rector, A. (2007). Semantic mapping of clinical model data to biomedical terminologies to facilitate data interoperability. In *HealthCare Computing Conference*.

Rector, A. (1999). Clinical terminology : why is it so hard ? *Methods Informatis in Medecine*, 38(4/5) :239–252.

Rector, A. et al. (2001). The interface between information, terminology, and inference models. *Studies in health technology and informatics*, 1 :246–250.

Rector, A., Nowlan, W., Kay, S., et al. (1991). Foundations for electronic medical records. *Methods of information in medicine*, 30(3) :179–86.

Rector, A., Nowlan, W., Kay, S., Goble, C., Howkins, T., et al. (1993). A framework for modelling the electronic medical record. *Methods of Information in Medicine*, 32(2) :109.

Reix, R. (2004). *SI et management des organisations*. 5ıème édition. Vuibert.

Renaud-Salis, J., Lagouarde, P., and Darmoni, S. (2010). Etude des systèmes d'aide à la décision médicale. Rapport technique, Haute Autorité de Santé.

Riché, P. (1979). *Les écoles et l'enseignement dans l'Occident chrétien de la fin du V siècle au milieu du XI siècle*. Picard.

Rigg, J., of Pure, I. U., and on Quantities & Units in Clinical Chemistry, A. C. C. (1995). *Compendium of Terminology and Nomenclature of Properties in Clinical Laboratory Sciences : Recommendations 1995*. Blackwell Science.

Rogers, J., Puleston, C., and Rector, A. (2006). The clef chronicle : patient histories derived from electronic health records. In *Proceedings of the 22nd International Conference on Data Engineering Workshops*, pages x109–x109. IEEE.

Rokach, L., Maimon, O., and Averbuch, M. (2004). Information retrieval system for medical narrative reports. *Flexible Query Answering Systems*, pages 217–228.

Rosenbloom, S., Miller, R., Johnson, K., Elkin, P., and Brown, S. (2006). Interface terminologies : Facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association*, 13(3) :277–288.

Sakji, S. (2010). *Recherche d'information et indexation automatique des médicaments à l'aide de plusieurs terminologies de santé*. Thèse de doctorat, Université de Rouen.

Sakji, S., Elkin, P., and Darmoni, S. (2011). Évaluation de l'indexation des comptes rendus médicaux à l'aide d'un outil états-unien adapté pour le français. In *Systèmes d'information pour l'amélioration de la qualité en santé. Comptes rendus des quatorzièmes Journées francophones d'informatique médicale (JFIM).*, Informatique et Santé, pages 155–164, Tunis. Springer.

Sakji, S., Letord, C., Dahamna, B., Kergourlay, I., Pereira, S., Joubert, M., and Darmoni, S. (2009a). Automatic indexing in a drug information portal. *Studies in Health Technology and Informatics*, 148 :112–122. PSIP.

Sakji, S., Letord, C., Pereira, S., Dahamna, B., Joubert, M., and Darmoni, S. (2009b). Drug information portal in europe : information retrieval with multiple health terminologies. *Studies in Health Technology and Informatics*, 150 :497–501. PSIP.

Sanderson, H., Adams, T., Budden, M., and Hoare, C. (2004). Lessons from the central hampshire electronic health record pilot project : evaluation of the electronic health record for supporting patient care and secondary analysis. *British Medical Journal*, 328(7444) :875–878.

Scherrer, J. and Spahni, S. (1999). Healthcare information system architecture (hisa) and its middleware models. In *Proceedings of the AMIA Annual Symposium*, page 935. AMIA.

Schloeffel, P., Beale, T., Hayworth, G., Heard, S., and Leslie, H. (2006). The relationship between cen 13606, hl7, and openehr. In *Proceedings of the Health Informatics*

Conference (HIC) and the Health Informatics New Zeland (HINZ)), volume 7, pages 24–28. Health Informatics Society of Australia.

Schulz, S., Daumke, P., Fischer, P., and Müller, M. (2008). Evaluation of a document search engine in a clinical department system. In *Proceedings of the AMIA Annual Symposium*, volume 2008, page 647. AMIA.

Schumacher, R., Berkowitz, L., Abramson, P., and Liebovitz, D. (2010). Electronic health records : Physician's perspective on usability. *Human Factors*, 54(12) :816–820.

Silberzahn, N. (1997). *Le dossier médical informatisé : modélisation et consultation*. Thèse de doctorat, Université de Caen. UFR de Médecine.

Skrbo, A., Begović, B., Skrbo, S., et al. (2004). [classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes]. *Medicinski arhiv*, 58(1 Suppl 2) :138.

Smith, B. and Ceusters, W. (2006). Hl7 rim : An incoherent standard. *Health Technology and Informatics*, 124 :133–138.

Soualmia, L. (2004). *Etude et Evaluation d'Approches Multiples d'Expansion de Requêtes pour une Recherche d'Information Intelligente : Application au Domaine de la Santé sur Internet*. Thèse de doctorat, INSA de Rouen.

Soualmia, L., Dahamna, B., Thirion, B., and Darmoni, S. (2006). Strategies for health information retrieval. *Stud Health Technol Inform*, 124 :595–600.

Sournia, J. (1991). *Histoire de la médecine et des médecins*. Editions Larousse.

Sowa, J. (1983). *Conceptual structures : information processing in mind and machine*. Addison-Wesley.

Spackman, K., Campbell, K., CÃ, R., et al. (1997). Snomed rt : a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. AMIA.

Spat, S. (2007). *Prototype of a Medical Information Retrieval System (MIRS)*. Thèse de doctorat, Graz University of Technology.

Takai-Igarashi, T., Ryo, A., Kenji, S., Takahisa, F., Y., M., et al. (2011). On experiences of i2b2 (informatics for integrating biology and the bedside) database with japanese clinical patients data. *Bioinformation*, 2(6) :86–90.

Tamine, L., Zemirli, N., Bahsoun, W., et al. (2007). Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information. *Information-Interaction-Intelligence (I3)*, 7(1) :5–25.

Terry, A. L., Chevendra, V., Thind, A., Stewart, M., Marshall, J. N., and Cejic, S. (2010). Using your electronic medical record for research : a primer for avoiding pitfalls. *Family Practice*, 27(1) :121–126.

Thirion, B., Douyère, M., Soualmia, L., Dahamna, B., Leroy, J., and Darmoni, S. (2004). Metadata element sets in the cismef quality-controlled health gateway. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC'04)*, pages 1–12.

Thirion, B., Loosli, G., Douyère, M., and Darmoni, S. (2003). Metadata element set in a quality-controlled subject gateway : a step to a health semantic web. *Studies in Health Technology and Informatics*, 95 :707–712.

Thomas, H. P., Monica, P., Robert, K., and Dom, R. (2006). The transition to electronic documentation on a teaching hospital medical service. In *Proceedings of the AMIA Annuel Symposium*, volume 2006, pages 629–633. AMIA.

Van Bemmel, J. and Musen, M. (1997). *Handbook of Medical Informatics*. Springer-Verlag.

Weed, L. (1971). *Medical Records, Medical Education, and Patient Care : The Problem-Oriented Record as a Basic Tool*. Year Book Medical Publishers.

Wong, W., Liu, W., and Bennamoun, M. (2010). An ontology-based interface for improving information exploration. In *Proceedings of the IUI Workshop on Intelligent Visual Interfaces for Text Analysis (IVITA)*, pages 29–32.

Yousefi, A., Mastouri, N., and Sartipi, K. (2009). Scenario-oriented information extraction from electronic health records. In *Proceedings of 22nd IEEE International Symposium on the Computer-Based Medical Systems (CBMS'09)*, pages 1–5. IEEE.

Yu, W. and Yilayavilli, S. (2009). A semantic-based dynamic search engine design and implementation for electronic medical records. In *Proceedings of the 11th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 187–189. IEEE.

Zapletal, E., Rodona, N., Grabarab, N., and Degoulet, P. (2010). Methodology of integration of a clinical data warehouse with a clinical information system : the hegp case. *Studies in Health Technology and Informatics*, 16(Pt 1) :193–7.

Zeng, Q. and Cimino, J. (1997). Linking a clinical system to heterogeneous information resources. In *Proceeding of the AMIA Annual Fall Symposium*, pages 553–557. AMIA.

Zeng, Q. and Cimino, J. (2000). Providing multiple views to meet physician information needs. In *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences*, pages 9–pp. IEEE.

Zeng, Q., Cimino, J., and Zou, K. (2002). Providing concepts-views for clinical data using a knowledge-based system : an evaluation. *Journal of the American Medical Informatics Association*, 9(3) :294–305.

Zhou, L. and Hripcsak, G. (2007). Temporal reasoning with medical data–a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40.

Zillner, S., Hauer, T., Rogulin, D., Tsymbal, A., Huber, M., and Solomonides, T. (2008). Semantic visualization of patient information. In *Proceedings of the 21st IEEE International Symposium on Computer-Based Medical Systems (CBMS'08)*, pages 296–301. IEEE Computer Society.

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

Troisième partie

ANNEXES

ANNEXE A

Données d'évaluation

A.1 Scénarios cliniques RIDoPI

A.1.1 Listes des cas cliniques

Cas clinique 3 :
M X, 62 ans, hypertendu, est admis aux urgences pour des troubles du langage. Rapidement, l'interne de garde suspecte un accident vasculaire cérébral et prescrit une IRM. Celle ci, réalisée deux heures après le début des symptômes, révèle des signes d'ischémie récente. Avant d'effectuer une thrombolyse, le neurologue doit rechercher d'éventuelles contres indications à cet acte thérapeutique, dans les plus bref délais [1]. En cas d'impossibilité de communiquer avec le patient, il pourra lancer l'outil de RI dans le DMP pour chercher des facteurs de risque hemorragiques (varices oesophagiennes, anévrisme artériel, ulcère récent, ...).

Cas clinique 4 :
M H, patient VIH+ depuis 12 ans, sous tri-thérapie depuis 7 ans (récement passé sous ATRIPLA), est hospitalisé pour une altération de l'état général et une toux sèche. Une pneumocystose est diagnostiquée et traitée rapidement par BACTRIM. Sans surprise, la biologie met en évidence un rebond virologique (CD4 <200/mm3 et charge virale $\tilde{1}$0 000 copies/ml). Il apparaît rapidement que face aux effets indésirables de l'ATRIPLA, le pa-

1. Recommandation de bonne pratique : Accident vasculaire cérébral : prise en charge précoce (alerte, phase préhospitalière, phase hospitalière initiale, indications de la thrombolyse). Haute Autorité de Santé. Mai 2009

tient ne prenait plus régulièrement son traitement. Le médecin se demande alors quelles spécialités proposer à son patient. Les carences du système d'information lui imposant un long et fastidieux travail de recensement des différents traitements déjà essayés et les différents niveaux de réponse biologique et clinique.

Cas clinique 5 :
Mme D est suivit depuis dix ans pour trouble bipolaire par le docteur V. Elle est hospitalisée ce jour pour tentative de suicide. Elle a apparemment bien suivi son traitement qui n'est pas assez efficace. Pendant l'entretien, Dr V se demande combien de tentative de suicide cette patiente a réalisé, il se demande aussi quelle molécule lui prescrire, en évitant celles qui lui ont déjà été prescrites sans montrer d'efficacité notable. Ces informations sont contenues dans ses notes, mais le dossier, depuis dix ans, est trop épais pour qu'il puisse retrouver rapidement ces informations.

Cas clinique 6 :
Le docteur B, après avoir regardé de près les 5 derniers EEG qu'il a fait passer à des patients atteints d'épilepsies, est arrivé à la conclusion qu'on pouvait distinguer deux entités nosologiques distinctes. Pour confirmer son impression, il souhaiterait passer en revue les dossiers de 50 patients épileptiques qui ont bénéficiés d'EEG. Il lance donc l'outil RIDOPI et saisi la requête suivante : *Epilepsie[CIM-10] ET EEG[CCAM]* l'interpréteur doit comprendre : (G40 ou ses fils) et (1.1.1.4 ou ses fils) et ramener les bons patients.

A.1.2 Fiche résumé DPI et Questions cliniques

CAS n° 7	
ID Patient	8
Sexe	F
Age	58
Nb de prises en charge	168
Pathologies codées	Maladie mitrale - insuffisance aortique - valves mécaniques - hétérogreffe - Insuffisance cardiaque - flutter auriculaire HTAP Kyste de l'ovaire en 1999 Anémie par carence en G6PD - beta thalassémie - sphingolipidose Effets indésirable des sulfamides Patient sous AVK - accident des AVK Cystite 04/04/2006 - insuffisance rénale chronique Infection après acte - complication des actes médicaux Intoxication par les opioïdes
Pathologies non codées et exploitables	Nombreux codes CIM9 et Méhari
Question clinique 1	Rechercher la date d'apparition de l'insuffisance ventriculaire gauche
Question clinique 2	Rechercher si il y a eu des complications des anticoagulants (Y44.2) et si il y a eu un TQ-PCENT <20% dans les 10 jours précédents la date de codage

FIGURE A.1 – Fiche résumé du DPI du patient 6 lié la question clinique Q4

CAS n° 1	
ID Patient	1
Sexe	M
Age	41
Nb de prises en charge	93
Pathologies codées	Epilepsie Diabète Retard mental PAVP - trauma crânien - hémorragie - triventriculaire - fracture bimalléolaire janvier 2009 Infections cutanées en janvier 2010
Pathologies non codées et exploitables	Nombreux codes CIM9 et Méhari
Question clinique 1	Rechercher le dernier électrocardiogramme (ECG) précédant la destruction d'un foyer arythmogène atriale
Question clinique 2	Rechercher les épisodes infectieux

FIGURE A.2 – Fiche résumé du DPI du patient 1 lié la question clinique Q2

A.2 Scénarios cliniques I2B2

Req #	Description
1	PI is looking at fungal infections in MICU patients, and needs to know what antibiotics patients received and for what periods of time. Specifically interested in looking at administration of IV antibiotics in this cohort while the patients were intubated.
2	PI needs counts of patients that had Basal cell Carcinomas, Squamous Cell Carcinomas, Squamous Cell Carcinomas in situ, Displastic Nevus, any kind of Nevus, melanomas, etc. Broken down by years 2004 and 2005
3	PI would like to look at all stroke patients, and would like the length of stay with and without outliers (outlier = 3 x std dev).
4	PI would like a list of patients who had hip/knee surgery and had received Rifampicin after their surgery.
5	PI is interested in looking at patients who were intubated greater than 48 hrs and received Propofol or Ativan while being intubated.
6	Female patients from January 2002 onwards with a diagnosis of DVT/PE and age less than 51.
7	PI would like trending data for her clinic INR(International normalized ratio) with the number of monthly INRs, number of critical INRs, percentage of INR within therapeutic range, diagnoses & total number of patients in the clinic.
8	PI needs a list of patients who have undergone Orthopedic surgery with a therapeutic INR on the day of surgery.
9	PI needs to find d-dimer values, dates & times on patients who have undergone Neurosurgery.
10	PI has requested list of names, MRNs of patients receiving Methylprednisolone gtt for acute spinal cord injury, including dates and doses for the period 01/01/2001 to 01/01/2007
11	PI has requested data on patients with TTP, or who have undergone plasmapheresis, or who have had the test ADAMTS13 ordered on them. The data elements requested contain CBCs, Chemistry panels, PT/INR and other labs with nadir and peak values for each of the lab results.
12	List of patients who have had D-dimer tests ordered or duplex ultrasound studies performed, along with the test/procedure dates and test-results.
13	PI needs data on patients with a diagnosis of Hypertension or systolic blood pressure greater than 130 mm Hg. Data elements include comorbidities (Diabetes, CAD, Stroke, etc.), procedures (CABG, coronary angiography, etc.) age and other patient demographics.
14	PI wants a report of all admits for patients with asthma, COPD, emphysema, atopic dermatitis or hay fever.
15	A list of pediatric burn patients (defined as age less than 15 years, and ICD-9 diagnosis codes 940-949, including all the sub-codes), who received low-molecular weight Heparin (Enoxaparin) during their hospital-stay, and who had anti-Factor Xa levels measured.

FIGURE A.3 – Requêtes I2B2 (1-15) Deshmukh et al. [2009]

Req #	Description
16	This study will evaluate the use of Echocardiography in driving management decisions in critically ill burn patients. PI needs to identify all burn patients (ICD Codes 940-949) during the last 5 years who had an Echo procedure.
17	This is a pre-research data request for an estimate of sample size on patients with non-ruptured cerebral aneurysm (ICD Code 437.30) who have had at least two Head MRA procedures.
18	PI needs a list of all ICU patients with a separate list for SICU patients who had an admission weight greater than 120kg.
19	List of patients with STEMI / PCI from 4/1/2008 - 6/30/2008
20	Need total number of new patients with the following diagnosis (ICD-9) codes billed from January - June 2008 by three physicians. Diagnosis codes: 596.59, 596.51, 788.31, 788.63, 788.41, either as primary or secondary diagnoses.
21	PI needs an estimate of how many patients were diagnosed with Aortic Stenosis from the Cardiology 'Lynx' application database that is used in the non-invasive cardiology lab.
22	Number of cases of childhood epilepsy and how many of these link to third generation families in the UPDB. Since multiple cases occur in a single family, user has requested both datasets: number of cases from the data warehouse as well as from the UPDB.
23	User needs a count from the Enterprise Data Warehouse of the patients with stroke. If possible, she would like to also know how many of them link to the UPDB.
24	The number of interventional cases (any case with PTCA, stent, alcohol ablation, PFO closure, ASD closure). List the number of cases by year from 2006, 2007, and 2008
25	PI has requested the following data-elements: Main hospital MRNs, patient names, discharge dates, times, units, discharging physician, discharge disposition, discharged with prescriptions; Patient-care units: 6N
26	PI has requested the number of CABG cases alone or with other types of procedure per month since June 2006 to the present.
27	Patient level detail for all patients that have had cardiac surgery since 7/1/2008

FIGURE A.4 – Requêtes I2B2 (16-27) Deshmukh et al. [2009]

Request #	Patient counts	Actual data	Diagnoses	Procedures	Medications	Labs	Demographics
1	•	•			•		•
2	•		•				
3	•	•	•				•
4	•			•	•		
5	•	•		•	•		
6	•	•	•				•
7	•	•		•		•	
8	•	•		•		•	•
9	•	•		•		•	•
10	•	•	•		•		•
11	•	•	•	•		•	•
12	•	•		•		•	•
13	•	•	•	•		•	•
14	•	•	•		•		•
15	•	•	•		•	•	•
16	•	•	•	•			•
17	•		•	•			
18	•	•					•
19	•	•	•	•			•
20	•		•				
21	•		•	•			
22	•		•				•
23	•		•				•
24	•			•			
25	•	•			•		•
26	•			•			
27	•	•		•			•
Total	**27**	**18**	**15**	**15**	**7**	**7**	**18**
%	**100%**	**66.6%**	**55.5%**	**55.5%**	**25.9%**	**25.9%**	**66.6%**

FIGURE A.5 – Types de données associés aux requêtes Deshmukh et al. [2009]

Request #	No significant modifications	Modifications required							
		Institution specific	Pre-processing	Post-processing	Exception conditions	Temporal conditions	Calculated fields	Additional attributes	Metadata modification
1			•	•				•	•
2	•					•			
3			•	•			•	•	•
4				•		•			
5	•				•				
6	•								
7			•	•			•		•
8			•	•		•	•		•
9	•			•					
10	•			•					
11			•	•		•	•	•	•
12			•					•	•
13		•	•	•		•	•	•	•
14		•	•	•				•	•
15	•								
16	•								
17	•								
18			•				•	•	•
19	•					•			
20		•	•			•		•	•
21		•	•					•	•
22		•	•	•	•	•	•	•	•
23		•	•	•	•	•		•	•
24	•					•			
25		•	•	•		•	•	•	•
26	•			•		•	•		
27	•			•		•			
Total	12	7	14	15	3	13	9	12	14
%	44.4%	25.9%	51.8%	55.5%	11.1%	48.1%	33.3%	44.4%	51.8%

FIGURE A.6 – Types de traitements associés aux requêtes Deshmukh et al. [2009]

A.3 2 descriptions du graphe RDF représentant le concept F45.33

```
@prefix : <http://www.chu-rouen.fr/smts#> .
@prefix basicontology: <http://www.mondeca.com/system/basicontology#> .
@prefix itm: <http://www.mondeca.com/system/itm#> .
@prefix publishing: <http://www.mondeca.com/system/publishing#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix skosm: <http://www.w3.org/2004/02/skos/mapping#> .
@prefix t3: <http://www.mondeca.com/system/t3#> .

:CIM10_BTNT~16514     a publishing:BT-NT;
     publishing:BT <http://www.chu-rouen.fr/smts#CIM10_F45.3>;
     publishing:NT <http://www.chu-rouen.fr/smts#CIM10_F45.33> .

:CIM10_F45.33     a :CIM10subdivision;
     :CIM10sort "F4533";
     :CIM10type "D";
     publishing:Level "5";
     rdfs:label "Somatoform autonomic dysfunction | Respiratory system"@en,
                "dysfonctionnement neurovégétatif somatoforme | Système respiratoire"@fr;
     skos:notation "F45.33" .
```

FIGURE A.7 – Notation N3

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:itm="http://www.mondeca.com/system/itm#"
   xmlns:skos="http://www.w3.org/2004/02/skos/core#"
   xmlns:publishing="http://www.mondeca.com/system/publishing#"
   xmlns:skosm="http://www.w3.org/2004/02/skos/mapping#"
   xmlns:t3="http://www.mondeca.com/system/t3#"
   xmlns:smts="http://www.chu-rouen.fr/smts#"
   xmlns:basicontology="http://www.mondeca.com/system/basicontology#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">

 <smts:CIM10subdivision rdf:about="http://www.chu-rouen.fr/smts#CIM10_F45.33">
   <rdfs:label xml:lang="en">Somatoform autonomic dysfunction | Respiratory system</rdfs:label>
   <rdfs:label xml:lang="fr">dysfonctionnement neurovégétatif somatoforme | Système respiratoire</rdfs:label>
   <smts:CIM10type>D</smts:CIM10type>
   <publishing:Level>5</publishing:Level>
   <smts:CIM10sort>F4533</smts:CIM10sort>
   <skos:notation>F45.33</skos:notation>
 </smts:CIM10subdivision>

 <publishing:BT-NT rdf:about="http://www.chu-rouen.fr/smts#CIM10_BTNT~16514">
   <publishing:NT rdf:resource="http://www.chu-rouen.fr/smts#CIM10_F45.33"/>
   <publishing:BT rdf:resource="http://www.chu-rouen.fr/smts#CIM10_F45.3"/>
 </publishing:BT-NT>

</rdf:RDF>
```

FIGURE A.8 – Notation RDF/XML

A.4 SPARQL-Joseki

Oracle SPARQL Endpoint Query Results

p	o
<http://www.mondeca.com/system/publishing#Level>	"5"
<http://www.w3.org/1999/02/22-rdf-syntax-ns#type>	<http://www.chu-rouen.fr/smts#CIM10subdivision>
<http://www.chu-rouen.fr/smts#CIM10type>	"D"
<http://www.w3.org/2000/01/rdf-schema#label>	"Somatoform autonomic dysfunction | Respiratory system" @en
<http://www.w3.org/2000/01/rdf-schema#label>	"dysfonctionnement neurovégétatif somatoforme | système respiratoire" @fr
<http://www.chu-rouen.fr/smts#CIM10sort>	"F4533"
<http://www.w3.org/2004/02/skos/core#notation>	"F45.33"

FIGURE A.9 – Résultats de la requête dans Joseki : la description détaillée du concept CIM-10 (F45.33)

Oracle SPARQL Endpoint Query Resu

fils	label
<http://www.chu-rouen.fr/smts#CIM10_A00.0>	"à Vibrio cholerae 01, biovar cholerae" @fr
<http://www.chu-rouen.fr/smts#CIM10_A00.1>	"à Vibrio cholerae 01, biovar El Tor" @fr
<http://www.chu-rouen.fr/smts#CIM10_A00.9>	"choléra, sans précision" @fr
<http://www.chu-rouen.fr/smts#CIM10_A01.0>	"fièvre typhoïde" @fr
<http://www.chu-rouen.fr/smts#CIM10_A01.1>	"paratyphoïde A" @fr
<http://www.chu-rouen.fr/smts#CIM10_A01.2>	"paratyphoïde B" @fr
<http://www.chu-rouen.fr/smts#CIM10_A01.3>	"paratyphoïde C" @fr
<http://www.chu-rouen.fr/smts#CIM10_A01.4>	"paratyphoïde, sans précision" @fr
<http://www.chu-rouen.fr/smts#CIM10_A02.0>	"entérite à Salmonella" @fr
<http://www.chu-rouen.fr/smts#CIM10_A02.1>	"septicémie à Salmonella" @fr
<http://www.chu-rouen.fr/smts#CIM10_A02.2>	"infection localisée à Salmonella" @fr
<http://www.chu-rouen.fr/smts#CIM10_A02.8>	"autres infections précisées à Salmonella" @fr

FIGURE A.10 – Extraits des résultats de la requête dans Joseki : trouver les concepts CIM-10 utilisés pour coder un épisode infectieux

ANNEXE B

Modélisation

B.1 Modèles

B.1.1 Métadonnées de notre modèle

B.1.2 Méta modèle CISMeF

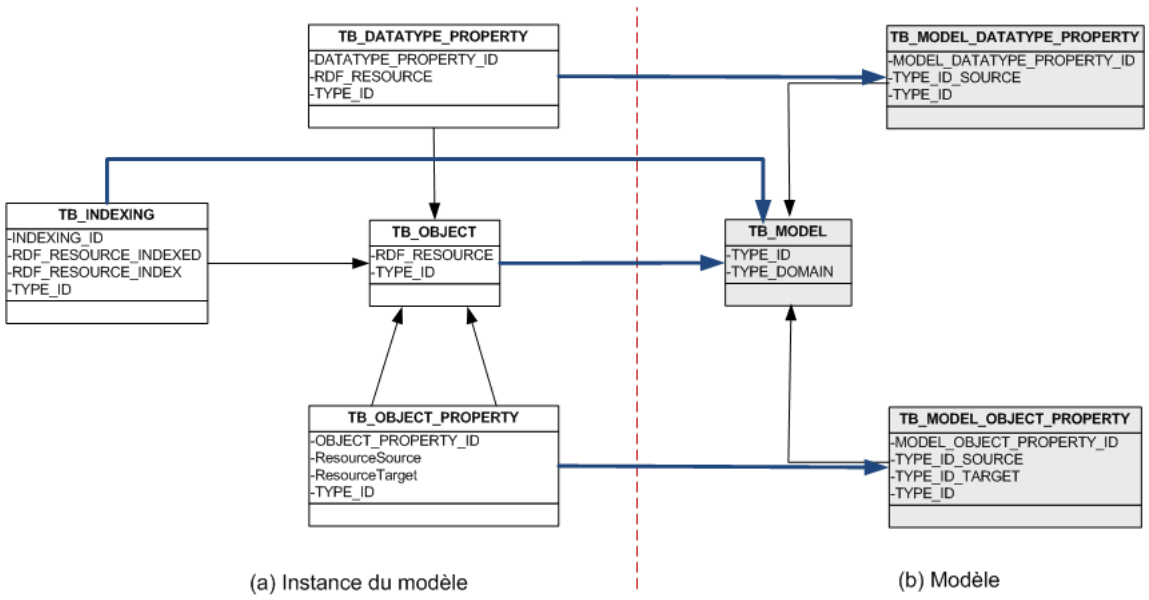

FIGURE B.1 – Méta modèle CISMeF simplifié

Description	Modèle EI@DM	Méta modèle CISMeF
Elements informationnels (EI)		
Patient	T_INFO_PAT	DM_PAT
Séjour	T_INFO_SEJ	DM_STAY
Acte	T_INFO_ACT	DM_ACT
Analyse biologique	T_INFO_ANABIO	DM_ANABIO
Compte-rendu (CR)	TT_INFO_CR	DM_REC
Bloc de CR	T_INFO_CRBLOC	
Attributs		
PATIENT		
Sexe	T_ATTR_SEXE	DM_SEX_PAT
Identifiant Unique Patient (NIP)	T_ATTR_IUP	DM_IUP_PAT
Année Naissance	T_ATTR_ANNEE	DM_ANNEE_PAT
Identifiant Patient (base CDP)	T_ATTR_IDPATIENT	DM_ID_PAT
Commune	T_ATTR_CODECOMMUNE	DM_COMMUNE_PAT
SEJOUR		
Identifiant Séjour (base CDP)	T_ATTR_IPECUFRM	DM_ID_STAY
Numéro Dossier	T_ATTR_NIP	DM_NIP_STAY
Type de séjour	T_ATTR_TYPESEJ	DM_MOD_STAY
Modalité de séjour (entrée et sortie)	T_ATTR_MOD_ENTREE et T_ATTR_MOD_SORTIE	DM_MOD_IN et DM_MOD_OUT
Date de séjour (entrée et sortie)	T_ATTR_DATE_ENTREE et T_ATTR_DATE_SORTIE	DM_DATE_IN et DM_DATE_OUT
Unité médicale (service)	T_ATTR_UM	DM_UM
ACTE		
Identifiant Acte (base CDP)	T_ATTR_IPECUFRM	DM_ID_STAY
Numéro Dossier	T_ATTR_NIP	DM_NIP_STAY
Date d'acte	T_ATTR_DATE_ACTE	DM_DATE_ACTE
Unité médicale (service)	T_ATTR_UM	DM_UM

TABLE B.1 – Métadonnées (partie 1)

B.1.3 Transposition de notre modèle vers le modèle CISMeF

Description	Modèle EI@DM	Méta modèle CISMeF
ANALYSE BIOLOGIQUE		
Identifiant Examen Biologique (base CDP)	T_ATTR_IDBIOEXAM	DM_IDBIO_EXAM_ANA
Identifiant Résultat Biologique	T_ATTR_IDANABIO	DM_ID_ANABIO
Valeur Analyse Biologique	T_ATTR_RES	DM_VAL_ANABIO
Norme Analyse Biologique	T_ATTR_NORME	DM_NORME_ANABIO
Borne Min (norme)	T_ATTR_NORMEMIN	DM_NORMMIN_ANABIO
Borne Max (norme)	T_ATTR_NORMEMIN	DM_NORMMIN_ANABIO
Paramètre Analyse	T_ATTR_PARAMANA	DM_PARAMANA_ANABIO
COMPTE-RENDU		
Identifiant Courrier (base CDP)	T_ATTR_IDCOURRIER	DM_ID_COURRIER
Type Compte-Rendu	T_ATTR_TYPECR	DM_TYPECR
Type Bloc CR	T_ATTR_TYPE_BLOCR	DM_TYPE_BLOCR
Relation_EIs		
Relation Séjour-Patient	T_REL_SEJPAT	DM_HAS_PAT(*)
Relation Séjour-Acte	T_ACT_SEJACT	DM_HAS_ACT(*)
Relation Séjour-Analyse	T_REL_SEJANABIO	DM_HAS_ANABIO(*)
Relation Séjour-CompteRendu	T_REL_SEJCR	DM_HAS_REC(*)
Relation Acte-CompteRendu	TT_REL_ACTCR	DM_HAS_REC(*)
Relation CompteRendu-Bloc	T_REL_CRBLOCCR	DM_HAS_REC_BLOC(*)

TABLE B.2 – Métadonnées (partie 2)

MODELE

TB_MODEL

TYPE_ID	XSD_TYPE	TYPE_DOMAINE	ANNOTATION
DM_STAY		OBJECT	CDP_TABLE "SEJOUR"
DM_UM		OBJECT	CDP_COLUMN "SEJOUR.UM_SEJ
DM_ID_STAY		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.IDPECUF_RM"
DM_MOD_STAY		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.TYPE_SEJ"
DM_MOD_IN		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.MODE_ENTREE"
DM_MOD_OUT		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.MODE_SORTIE"
DM_DATE_INT		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.DATEENTREE"
DM_DATE_OUT		DATATYPE_PROPERTY	CDP_COLUMN "SEJOUR.DATESORTIE"
DM_IS_STAY		MODELE_OBJECT_PROPERTY	RELATION "EST_STAY"
DM_HAS_STAY		MODELE_OBJECT_PROPERTY	RELATION "A_STAY"
DM_IS_UM		MODELE_OBJECT_PROPERTY	RELATION "EST_UM"
DM_HAS_UM		MODELE_OBJECT_PROPERTY	RELATION "A_UM"

TB_MODEL_OBJECT_PROPERTY

TYPE_ID	TYPE_ID_SOURCE	TYPE_ID_TARGET
HAS_ATTRIBUT	DM_STAY	DM_ID_STAY
HAS_ATTRIBUT	DM_STAY	DM_NIP
HAS_ATTRIBUT	DM_STAY	DM_MOD_STAY
HAS_ATTRIBUT	DM_STAY	DM_MOD_IN
HAS_ATTRIBUT	DM_STAY	DM_MOD_OUT
HAS_ATTRIBUT	DM_STAY	DM_DATE_INT
HAS_ATTRIBUT	DM_STAY	DM_DATE_OUT
HAS_ATTRIBUT	DM_STAY	DM_UM_STAY
HAS_RELATION	DM_STAY	DM_HAS_PAT
HAS_RELATION	DM_STAY	DM_IS_STAY
BIJECTION	DM_HAS_STAY	DM_IS_STAY
HAS_RELATION	DM_STAY	HAS_UM
HAS_RELATION	DM_UM	DM_IS_UM
BIJECTION	DM_HAS_UM	DM_IS_UM

FIGURE B.2 – Objet "SEJOUR" dans le modèle CISMeF

INSTANCE MODELE

TB_OBJECT

RDF_RESOURCE	TYPE_ID
SEJ_114005	DM_STAY

TB_OBJECT_PROPERTY

TYPE_ID	RDF_RESOURCE_SOURCE	RDF_RESOURCE_TARGET
DM_HAS_PAT	SEJ_114005	PAT_1095293
DM_HAS_UM	SEJ_114005	UM_MSEJ

TB_DATATYPE_PROPERTY

RDF_RESOURCE	TYPE_ID	VALUE
SEJ_114005	DM_ID_STAY	SEJ_114005
SEJ_114005	DM_NIP	584329210
SEJ_114005	DM_MOD_STAY	H
SEJ_114005	DM_MOD_IN	Domicile
SEJ_114005	DM_MOD_OUT	Domicile
SEJ_114005	DM_DATE_INT	06/04/07
SEJ_114005	DM_DATE_OUT	18/05/07

TB_INDEXING

RDF_RESOURCE_INDEXED	RDF_RESOURCE_INDEX	TYPE_ID	ENTRY_TYPE
SEJ_114005	CIM10_G40.0	INDEX	ET_MANU
SEJ_114005	CIM10_F73.9	INDEX	ET_MANU

FIGURE B.3 – Instances de l'objet "SEJOUR" dans le modèle CISMeF

B.2 Transformation d'un DPI en documents CDA HL7

```
<!-- ********************************************************************************************************
Date de création 06/09/2010
Format : conforme ANSI/HL7 CDA, R2-2005 4/21/2005
Objectif : Exemple d'un document HL7/CDA pour le patient N°1095293
********************************************************************************************************-->
<?xml-stylesheet type="text/xsl" href="CDA_fr.xsl"?>
<ClinicalDocument
    xmlns="urn:hl7-org:v3"
    xmlns:voc="urn:hl7-org:v3/voc"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="urn:hl7-org:v3 CDA.xsd">
    <!--  ********************************************************
      HEADER - HL7 France
    ********************************************************    -->
    <realmCode code="FR"/>
    <typeId root="2.16.840.1.113883.1.3" extension="POCD_HD000030" />
    <templateId root="2.16.840.1.113883.3.27.1776" />
    <!--
        Identifiant Unique du document, possibilité d'extension de l'identifiant
    -->
    <id root="PAT_1095293" extension="NIP_1072606091"  />
    <!--
        Code du document qui résumes les épisodes (séjour, acte, examen biologique, etc.) de soins d'un patient
     -->
    <code   code="34133-9" codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"
                displayName="SUMMARIZATION OF EPISODE NOTE" />

    <title>Episodes de soins - Patient N°1055293</title>
    <effectiveTime value="201009061300+0200"/>
    <confidentialityCode code="N" codeSystem="2.16.840.1.113883.5.25" />
    <languageCode code="fr-FR" />
    <versionNumber value="2" />
    <!--  ********************************************************
      REFERENCES DOSSIER PATIENT
    ********************************************************    -->
    <recordTarget>
        <patientRole>
            <!--
                Dépend de l'Identifiant National de Santé (INS)
                root : OID de l'Autorité de l'affaction du N° INS
             -->
            <id extension="1095293" root="1.2.250.1.143.1" />

            <patient>
                <administrativeGenderCode code="F" codeSystem="2.16.840.1.113883.5.1" />
                <birthTime value="19180101" />
            </patient>
        </patientRole>
```

FIGURE B.4 – Extrait (en-tête) des épisodes de soins d'un patient au format HL7

Rouen le 18/04/2006
COMPTE-RENDU D'HOSPITALISATION :

XX XX - Né(e) le : jj mm/1954

Date d'entrée : 17/04/2006 N° dossier
Date de sortie : 17/04/2006

Réf. : ED/LL

Compte-rendu de biothérapie

* Motif d'hospitalisation

Evaluation à 6 mois de l'efficacité et de la tolérance d'un traitement par HUMIRA dans le cadre d'une maladie de Crohn, associée à un psoriasis cutané, toutefois sans manifestation rhumatismale.

*Antécédents

Intercure
Infections : un épisode de rhino-pharyngite en janvier ne justifiant pas l'arrêt de l'HUMIRA.
Disparition de toute plaque de psoriasis.
Absence de problème articulaire.
Autres : Madame XX conserve quelques douleurs pelviennes, 6 à 7 selles par jour, parfois nocturnes.

* Examen physique

Poids : 62 kg. TA : 14/9. Température : 37,4°.
Schöber : 14,5. Distance mains-sol : 27. Ampliation thoracique : 4. Distance occiput-mur : 5.
Sacro-iliaques indolores.
Auscultation cardio-pulmonaire, examen ORL (à gauche petit érythème du canal auditif externe),
Examen des organes hématopoïétiques : normaux.
Absence de psoriasis.
Absence de douleur articulaire.

* Examens complémentaires:

BU : sang et leucocytes.
Autres : ECBU : GB entre 100 et 1000/mm3 ; GR lyses, flore polymicrobienne.
VS : 26. CRP < 3. NFS, plaquettes : normales avec VGM à 89 normal.
FAN : 1/1200. Ac anti Ro : négatif. Ac anti DNA : négatif.

* CONCLUSION

Excellente tolérance de l'HUMIRA et excellente efficacité sur le plan du psoriasis cutané.
Amélioration des symptômes de la maladie de Crohn.
Augmentation de la positivité des FAN.
Sur le plan urinaire, possible contamination lors du recueil. Contrôle ECBU à renouveler.

* Conduite à tenir / Rendez-vous
Diminution du traitement de SPECIAFOLDINE pris à raison de 2 cp par jour passé à 2 cp par semaine en l'absence de macrocytos
A revoir dans trois mois pour surveillance de la prise d'HUMIRA en l'absence de consultation prévue avec le Professeur LEREB
Doit rencontrer son dentistes pour surveillance le jeudi 6 avril.
Courrier adressé à dentiste pour lui signaler la conduite à tenir en cas de soins sanglants.
On rappelle à Madame XX les règles de suspension de l'HUMIRA lors de la survenue de toute infection.
Surveillance biologique mensuelle.

* Traitement de sortie

METHOTREXATE	20 mg/sem le vendredi
HUMIRA	40 mg tous les 2 jeudis par infirmière
SPECIAFOLDINE	2 cp le dimanche

FIGURE B.5 – Compte-rendu d'hospitalisation n°4779106

```
            MEDICATIONS
            Conclusion :
                - A revoir dans trois mois pour surveillance de la prise d'HUMIRA
            Traitement :
                - HUMIRA            40 mg tous les 2 jeudi
                - SPECIAFOLDINE     2 cp le dimanche
*****************************************************************************************-->
<component>
    <section>
        <code code='10183-2' displayName='HOSPITAL DISCHARGE MEDICATIONS' codeSystem='2.16.840.1.113883.6.1' codeSyste
        <entry>
            <substanceAdministration classCode='SBADM' moodCode='INT'>
                <statusCode code='completed'/>
                <effectiveTime xsi:type='IVL_TS'>
                <low value='20100101'/>
                <high value='20100401'/>
                </effectiveTime>
                <effectiveTime xsi:type="PIVL_TS" operator="A">
                    <period value="7" unit="d"/><!-- Chaque dimanche  -->
                </effectiveTime>
                <doseQuantity>
                    <center value="2"/>
                </doseQuantity>
                <consumable>
                    <manufacturedProduct>
                        <manufacturedLabeledDrug>
                            <code code="3040891" codeSystem="XX" codeSystemName="CIP" displayName="SPECIAFOLDINE 5 mg
                            <originalText mediaType="text/xml"><reference value="#med-1"/></originalText>
                        </manufacturedLabeledDrug>
                    </manufacturedProduct>
                </consumable>

            </substanceAdministration>
        </entry>
        <entry>
            <substanceAdministration classCode='SBADM' moodCode='INT'>
                <statusCode code='completed'/>
                <effectiveTime xsi:type='IVL_TS'>
                <low value='20100101'/>
                <width value="3" unit="m"/>
                </effectiveTime>
                <effectiveTime xsi:type="PIVL_TS" operator="A">
                    <period value="15" unit="d"/><!--  Tous les 2 jeudis  -->
                </effectiveTime>
                <doseQuantity>
                    <center value="1"/>
                </doseQuantity>
                <consumable>
                    <manufacturedProduct>
                        <manufacturedLabeledDrug>
                            <code code="ATCL04AB04" codeSystem="XX" codeSystemName="ATC" displayName="HUMIRA 40mg sol
                            <originalText mediaType="text/xml"><reference value="#med-1"/></originalText>
                            <translation code="3780145" codeSystem="XX" codeSystemName="CIP"/>
                        </manufacturedLabeledDrug>
```

FIGURE B.6 – Body CDA (partie 1) correspondant au CR médical n°4779106

```
            MEDICATIONS
            Conclusion :
                - A revoir dans trois mois pour surveillance de la prise d'HUMIRA
            Traitement :
                - HUMIRA            40 mg tous les 2 jeudi
                - SPECIAFOLDINE     2 cp le dimanche
*****************************************************************************************-->
<component>
    <section>
        <code code='10183-2' displayName='HOSPITAL DISCHARGE MEDICATIONS' codeSystem='2.16.840.1.113883.6.1' codeSyste
        <entry>
            <substanceAdministration classCode='SBADM' moodCode='INT'>
                <statusCode code='completed'/>
                <effectiveTime xsi:type='IVL_TS'>
                <low value='20100101'/>
                <high value='20100401'/>
                </effectiveTime>
                <effectiveTime xsi:type="PIVL_TS" operator="A">
                    <period value="7" unit="d"/><!-- Chaque dimanche  -->
                </effectiveTime>
                <doseQuantity>
                    <center value="2"/>
                </doseQuantity>
                <consumable>
                    <manufacturedProduct>
                        <manufacturedLabeledDrug>
                            <code code="3040891" codeSystem="XX" codeSystemName="CIP" displayName="SPECIAFOLDINE 5 mg
                            <originalText mediaType="text/xml"><reference value="#med-1"/></originalText>
                        </manufacturedLabeledDrug>
                    </manufacturedProduct>
                </consumable>

            </substanceAdministration>
        </entry>
        <entry>
            <substanceAdministration classCode='SBADM' moodCode='INT'>
                <statusCode code='completed'/>
                <effectiveTime xsi:type='IVL_TS'>
                <low value='20100101'/>
                <width value="3" unit="m"/>
                </effectiveTime>
                <effectiveTime xsi:type="PIVL_TS" operator="A">
                    <period value="15" unit="d"/><!--  Tous les 2 jeudis  -->
                </effectiveTime>
                <doseQuantity>
                    <center value="1"/>
                </doseQuantity>
                <consumable>
                    <manufacturedProduct>
                        <manufacturedLabeledDrug>
                            <code code="ATCL04AB04" codeSystem="XX" codeSystemName="ATC" displayName="HUMIRA 40mg sol
                            <originalText mediaType="text/xml"><reference value="#med-1"/></originalText>
                            <translation code="3780145" codeSystem="XX" codeSystemName="CIP"/>
                        </manufacturedLabeledDrug>
```

FIGURE B.7 – Body CDA (partie 2) correspondant au CR médical n°4779106