

Partage de données biomédicales : modèles, sémantique et qualité

Rémy Choquet

▶ **To cite this version:**

Rémy Choquet. Partage de données biomédicales : modèles, sémantique et qualité. Bio-Informatique, Biologie Systémique [q-bio.QM]. Université Pierre et Marie Curie - Paris VI, 2011. Français. NNT : 2011PA066467 . tel-00824931

HAL Id: tel-00824931

https://theses.hal.science/tel-00824931

Submitted on 22 May 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PIERRE ET MARIE CURIE

ECOLE DOCTORALE EPIDEMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMEDICALE

THESE

pour obtenir le titre de

Docteur en Sciences

de l'Université Pierre et Marie Curie, Paris 6

Mention : Informatique

Présentée et soutenue par

Rémy Choquet

Partage de données biomédicales : modèles, sémantique et qualité.

Thèse dirigée par Marie-Christine Jaulent

et Omar Boussaïd

préparée au laboratoire d'ingénierie des connaissances en santé

UMR872 EQ20, Projet DebugIT

soutenue le 16 Décembre 2011

Rapporteurs :	Régis Beuscart, PUPH	-	Université de Lille 2
	Chantal Reynaud, PU	-	Université Paris-Sud
Examinateurs :	Jean-Gabriel Ganascia, PU	-	Université Paris 6
	Dirk Colaert, MD	-	Agfa Healthcare
Directeurs :	Marie-Christine Jaulent, DR	-	UMR_S 872 Éq.20
	Omar Boussaïd, PU	-	Université de Lyon 2

"Suggérer, c'est créer. Décrire, c'est détruire."
Robert Doisneau, photographe.

Remerciements

Marie-Christine Jaulent, ma directrice de thèse, qui m'a poussé à mettre de l'ordre dans mes pensées en m'apprenant la pensée scientifique. Marie-Christine m'a ouvert au monde de l'information médicale. Je ne pourrais jamais assez la remercier de la confiance qu'elle a mis en moi, de l'indéfectible support, et du temps qu'elle a toujours trouvé pour échanger. Marie-Christine fait partie de ces rares et précieuses personnes qui vous aident à aller où vous voulez aller, sans se soucier d'elle-même.

Omar Boussaïd, mon directeur de thèse, qui m'a emmené sur la voie de la recherche il y a quelques années, à Lyon, alors que je faisais une parenthèse sur mon parcours dans le monde professionnel. Omar m'a principalement appris à justement nommer et utiliser les concepts des théories de l'information. Et il m'a fait confiance.

Jos de Roo, logicien. Jos m'a aidé à comprendre qu'on ne peut imposer un modèle au monde, car la nature même de ce modèle est personnelle et c'est sa vocation. Construire des systèmes communicants c'est d'abord rendre explicite ce qui est implicite. Quand on aura atteint un niveau acceptable d'explicité, alors nous pourrons commencer à penser à faire communiquer les systèmes d'information. Tout comme l'homme, la machine doit être capable d'interpréter. Il faut lui en donner les moyens et en connaître les limites. L'espace d'interprétation est sûrement une faculté humaine liée à la création.

Dirk Colaert et Christian Lovis, principaux instigateurs du projet DebugIT. Deux hommes qui poussent la recherche médicale en Europe. Leur énergie a été vitale au projet DebugIT. Ils ont tous deux toujours eut l'écoute et le bon mot pour guider mes réflexions. Ils partagent, avec générosité.

L'équipe de développement d'AGFA (Giovanni Mels, Kristof Depraetere et Boris Devloed) qui ont accepté de travailler avec moi sur la plateforme d'interopérabilité sémantique du projet DebugIT. Cette thèse n'aurait simplement pas pu aboutir sans vous.

Jean Charlet, chercheur ontologue. Jean a su m'aider sur mes questionnements vis à vis des ontologies et de leur pouvoir d'expression. Jean m'a mis en garde, sur les sirènes des ontologies qui ne sont jamais utilisées réellement... Je remercie particulièrement Jean pour le temps qu'il a toujours trouvé pour échanger, et refaire le monde...

Et merci à toutes ces personnes qui ont croisé mon chemin au laboratoire SPIM/ICS, nous le savons tous, ce laboratoire est un exemple d'environnement chaleureux d'échange et de travail. J'ai maintenant la preuve qu'on peut allier passion, travail et amitié.

Table des matières

1 Introduction générale **3**
1.1 L'informatique médicale 4
1.1.1 Pourquoi une science de l'information médicale ? 4
1.1.2 L'évolution des architectures des systèmes d'information de santé 6
1.1.3 L'informatisation actuelle 6
1.2 Positionnement 7
1.3 Objectifs 8
1.4 Le contexte et l'expérimentation 9
1.5 Enjeux scientifiques 9

I Partage de données et connaissances biomédicales **11**

2 Modèles et Représentations **13**
2.1 Introduction 14
2.2 Modèles de données 16
2.2.1 Un bref historique 17
2.2.2 Le modèle relationnel 18
2.2.3 Le modèle de données objet 19
2.2.4 Le modèle dimensionnel 21
2.2.5 Le modèle de données dimensionnel semi-structuré 22
2.2.6 Le modèle EAV 24
2.2.7 Le modèle dimensionnel EAV 25
2.2.8 Le modèle associatif 25
2.2.9 Le modèle RDF 28
2.2.10 Synthèse 31
2.3 Modèles de connaissance 33
2.3.1 Définitions 34
2.3.2 Terminologies et Classifications 35
2.3.3 Ontologies 37
2.3.3.1 Anatomie d'une ontologie 38
2.3.3.2 Natures d'ontologies 41
2.4 Modèles de qualité 43

2.4.1 La qualité de l'information 43
2.4.1.1 Les mesures de qualité 43
2.4.1.2 Les processus d'amélioration de la qualité 44
2.4.1.3 Exemples de projets qualité 45
2.4.2 Synthèse 46
2.5 Synthèse 46
2.6 Discussion 47

3 Partage d'information 49
3.1 Introduction 50
3.2 Interopérabilité 51
3.3 Modèles d'intégration de données 53
3.3.1 Intégration centralisée et persistante 55
3.3.2 Intégration centralisée ou décentralisée non persistante 57
3.3.3 Intégration à la volée ou "mashup" 58
3.4 Le Web Sémantique 59
3.4.1 Langages de représentation 61
3.4.2 Langages de règles 63
3.4.2.1 SWRL / Semantic Web Rule Language 64
3.4.2.2 Turtle N3 64
3.4.3 SPARQL - Standard Protocol and RDF Query Language . . 65
3.4.4 Logiques et Raisonneurs 66
3.5 Conclusion 68

4 Standards en santé pour partager l'information 71
4.1 Introduction 72
4.1.1 Modèles : Health Level Seven (HL7) 72
4.1.2 Modèles : openEHR 74
4.1.2.1 Le modèle d'information d'openEHR 75
4.1.2.2 Les archétypes openEHR 75
4.1.2.3 Différences et complémentarités 77
4.1.3 Terminologies : NEWT / ATC / ICD10 78
4.1.4 Ontologie : SNOMED CT et UMLS 79
4.2 Interopérabilité Sémantique en Santé 80
4.2.1 Aggrégation de modèles, approche entrepôt de données 82
4.2.2 Alignement de modèles : fédération et médiation de données . 83
4.2.2.1 Intégration à la volée et "mashup" 84
4.3 Analyse Comparative 85

4.4 Synthèse et Discussion 86

II Partage de l'information biomédicale dans le domaine de l'émergence de la résistance aux antibiotiques 89

5 Agents infectieux et traitements antibiotiques 91
5.1 Introduction 92
5.2 Le contexte 92
5.2.1 L'infection bactérienne 93
5.2.1.1 Une relation comensale 93
5.2.1.2 L'infection nosocomiale 94
5.2.2 L'antibiothérapie 94
5.2.3 L'antibiogramme 95
5.2.4 L'évolution de la résistance 96
5.2.5 Les propositions pour contenir la pandémie 97
5.3 DebugIT 98
5.3.1 Introduction 98
5.3.2 Contributions attendues de DebugIT 99

6 Connaître et partager de l'information biomédicale 103
6.1 Introduction 104
6.2 Connaissances liées à la qualité de l'information 106
6.2.1 La qualité de l'information et la sémiotique 107
6.2.2 Qualité de l'information source pour l'interopérabilité 109
6.2.2.1 Evaluation 109
6.2.2.2 Alignement et Surveillance 110
6.3 Des données vers la sémantique 111
6.3.1 Interopérabilité Technique 112
6.3.2 Interopérabilité Syntaxique 113
6.3.2.1 Un modèle d'analyse standardisé 114
6.3.2.2 Une approche dimensionnelle enrichie par des ressources sémantiques 118
6.3.2.3 Normalisation des termes 119
6.3.3 Interopérabilité Sémantique 120
6.3.3.1 Data Definition Ontology 122
6.3.3.2 Métadonnées et Qualité 126
6.4 La plateforme d'interopérabilité sémantique 127

6.4.1 Introduction . . . 127
6.4.2 Fonctionalités générales d'IP . . . 128
6.4.2.1 DebugIT Core Ontology . . . 129
6.4.2.2 DebugIT terminologies . . . 130
6.4.2.3 Clinical Data Repository . . . 132
6.4.2.4 Fouille de données . . . 133
6.4.2.5 Aide à la décision . . . 133
6.4.3 Cas d'utilisation . . . 133
6.4.4 Réécriture de requêtes . . . 136
6.4.5 Le problème du monde ouvert . . . 140
6.4.6 Une proposition partiellement satisfaisante . . . 143
6.5 Conclusion . . . 144

7 Expérimentation 147
7.1 Introduction . . . 148
7.2 Évaluation de la qualité . . . 150
7.2.1 Mise en oeuvre . . . 150
7.2.1.1 Audit . . . 150
7.2.1.2 Qualification . . . 151
7.2.1.3 Normalisation et surveillance . . . 152
7.3 Constitution d'un entrepôt de données clinique . . . 154
7.3.1 Entrepôt de données de santé : Une modélisation standardisée 154
7.3.2 Les autres Clinical Data Repository européens . . . 155
7.4 La plateforme d'interopérabilité . . . 157
7.4.1 Formalisation d'une source de données . . . 157
7.4.2 La médiation sémantique de données . . . 159
7.4.2.1 Architecture générale . . . 159
7.4.2.2 Un exemple de médiation de données . . . 163
7.4.2.3 Résultats . . . 169
7.4.3 Validation de l'approche de médiation sémantique . . . 170
7.5 Conclusion . . . 172

III Conclusion Générale 175

8 Conclusions, discussions et perspectives 177
8.1 Introduction . . . 178
8.2 Partage et modèles . . . 178

8.3 Partage et qualité 180
8.4 Partage et sémantique 180
8.5 La plateforme d'interopérabilité sémantique 182
8.6 Contributions Personnelles 182
8.7 Conclusion générale 183

Bibliographie **187**

Table des figures

1.1 Interdisciplinarité de l'informatique médicale 5

2.1 Le problème de modélisation ontologique du mouton anglais 15
2.2 Le modèle relationnel 19
2.3 Le modèle de données objet 20
2.4 Le modèle dimensionnel 22
2.5 Une hiérarchie de dimension 23
2.6 Le modèle dimensionnel vertical 26
2.7 Une association simple 26
2.8 Une association sur une association 27
2.9 Un graphe décrivant Rémy Choquet 29
2.10 Une relation n-aire avec un identifiant universel pour l'adresse 30
2.11 Une relation n-aire avec un noeud vide. 30
2.12 Tableau récapitulatif des propriétés des structures de stockage d'information. 32
2.13 Le triangle aristotélicien (gauche) et le triangle sémiotique (droite) . 39
2.14 Exemple de relation de subsomption 40
2.15 Pyramide des niveaux ontologiques. Adaptés depuis Alan Rector. . . 41
2.16 Exemple d'ontologie de haut niveau, d'ontologie de domaine de haut niveau et d'ontologie de domaine. Adapté depuis Stefan Schulz. . . . 42
2.17 Le processus d'amélioration de la qualité Six Sigma. 45

3.1 La pile du web sémantique, première version. 53
3.2 3 approches d'intégration de données 54
3.3 L'architecture du projet WHIPS. 57
3.4 La pile du web sémantique actuelle 61
3.5 Données liées sur le web 62
3.6 Linked Data, Statut du Data Cloud en Septembre 2010. 70

4.1 Modèle conceptuel de référence d'HL7 : le RIM 73
4.2 Structure du modèle d'information de référence d'openEHR 76
4.3 Vue de l'architecture de caGrid. 84
4.4 Une vue des alertes d'épidémie au cours des 3 derniers mois précédent le 01/02/2011. 85

4.5 Tableau comparatif des projets d'intégration de données dans le domaine de la santé (excepté pour OntoDB). 86

5.1 Une bactérie mute et s'adapte à son milieu. 92
5.2 Les relations entre l'hôte, la bactérie et l'antibiotique dans le cas d'une infection avérée. (Image extraite d'un document du projet DebugIT) 95
5.3 Un exemple de résultat d'antibiogramme 96
5.4 La boucle DebugIT 101

6.1 Le triangle de qualité de l'information (TQI) 108
6.2 Vue logique de l'architecture de qualité de TransMED. 111
6.3 Les 3 couches de l'approche d'interopérabilité pour l'intégration de données de DebugIT 112
6.4 Une vue du flux de travail de mise en oeuvre du modèle multidimensionnel basé sur le domaine décrit dans HL7. 116
6.5 Le modèle de données physique relationnel basé sur les 6 modèles d'information HL7. 117
6.6 Les relations conceptuelles entre les faits et les dimensions du modèle dimensionnel 118
6.7 Une architecture d'entrepôt de données décisionnel sémantique hybride 119
6.8 Approches LAV ou GAV 121
6.9 Niveaux de représentation et alignements 122
6.10 3 niveaux de représentation de l'information 123
6.11 La plateforme d'interopérabilité doit être réutilisable avec d'autres données, et d'autres connaissances. 128
6.12 Représentation fonctionnelle d'IP 129
6.13 Extrait de la DCO. Représentation ontologique des concepts d'antibiothérapie. 131
6.14 L'architecture D2R : L'outil permet l'accès aux données relationnelles en format RDF en HTML, RDF et SPARQL. 133
6.15 Cas d'utilisation global. Représentation des acteurs de la plateforme d'interopérabilité. 134
6.16 Vue des interfaces d'IP et des ressources que la plateforme utilise pour opérer. 135
6.17 Vue logique du système de médiation de données dans IP. 137
6.18 Vue logique du système de médiation de données dans IP. 138
6.19 Processus d'enrichissement et de formalisation d'une source de données pour le partage et l'analyse de données 145

7.1 Notre expérimentation se situera à différents niveaux dans la problématique d'interopérabilité de données (en vert) 149
7.2 Résultat de la méthode d'évaluation subjective du modèle d'information du DPI de l'HEGP 152
7.3 Modèle dimensionnel HL7 pour l'étude de l'évolution de la résistance aux antibiotiques dans DebugIT 156
7.4 Vue conceptuelle de l'architecture générale d'IP, de ses services et de ses acteurs. 160
7.5 Vue globale des services de IP et de leur interactions 161
7.6 Diagramme de séquence de réécriture de requête. 162
7.7 Diagramme représentant le flux de tâches nécessaires à la création d'une nouvelle requête clinique dans DebugIT. 163
7.8 Diagramme d'exécution d'une requête clinique sur 2 CDRs 164
7.9 Gadget montrant un graphique de résistance de E.Coli aux Fluoroquinolones au endpoint de l'INSERM sur l'année 2007 ainsi que les données agrégées des autres sites. 170
7.10 Graphique de taux de sensibilité de E.Coli à la Trimethoprime à l'HEGP sur une période de 6 ans. 171
7.11 Graphique de taux de sensibilité de E.Coli à la Cefixime à l'HEGP sur une période de 6 ans. 172
7.12 Graphique de taux de sensibilité de E.Coli à la Chloramphenicol à l'HEGP sur une période de 6 ans. 173

A ma mère, à ma femme, à ma fille...

Chapitre 1

Introduction générale

"Un hôpital est une machine complexe générant de l'information" - [Grémy 1987]

L'information générée par un hôpital peut être une information extrêmement utile dans un contexte de soin lorsqu'elle est partagée, par exemple, par divers hôpitaux. A une échelle plus large que l'hôpital, l'analyse d'activité de soin peut permettre la mise en place de réseaux de surveillance ou bien simplement faciliter le suivi du soin des patients entre les professionnels de santé. Cette information générée n'est cependant pas aisément partageable à ce jour ni entre hôpitaux, ni à l'intérieur du même hôpital, ni entre l'hôpital et la médecine de ville. Plusieurs verrous sont identifiés dont le plus significatif est certainement la difficulté que l'on a à représenter la connaissance médicale de manière standardisée et unique, même à l'intérieur de la même spécialité médicale. Les structures de stockage (et par extension de partage) de l'information n'apportent pas non plus de réponse unique et concrète au problème de l'information biomédicale. Enfin, la qualité des données médicales reste un frein majeur à l'utilisation de celles-ci pour l'analyse médicale. Notre travail se situe de manière générale à la frontière de l'informatique et de la médecine. Il pose la question de l'utilisation des technologies de l'information et de la communication dans le cadre de l'amélioration des pratiques médicales, de la découverte de nouvelles connaissances ou bien de la recherche médicale. Nous tenterons de définir le concept de l'information appliqué à la machine en définissant les propriétés de ce concept et ses limites. Nous aborderons les notions de partage de cette information puis nous proposerons une approche de partage dans le cadre d'un projet européen. Plus généralement nous tenterons d'exposer clairement la place de la sémantique dans le paradigme du partage de l'information biomédicale à ce jour.

1.1 L'informatique médicale

La pratique d'une médecine moderne et de qualité ne peut être dissociée d'un traitement rationnel de l'information médicale. En effet, la complexité croissante de la médecine occidentale actuelle (spécialisation des médecins, quantité d'information à traiter, optimisation de la posologie des médicaments, guides de bonnes pratiques, etc.) poussent de manière naturelle à la mise en place de systèmes d'information capables d'aider le praticien dans ses tâches quotidiennes de prise en charge du patient. Pour cela, l'informatique médicale se nourrit des recherches issues de divers domaines comme l'ingénierie des connaissances, l'intelligence artificielle ou bien l'ingénierie des modèles. Chacun de ces domaines de recherche apportent à l'informatique médicale des méthodes, des techniques et des outils permettant d'améliorer la formalisation des données et des connaissances dans les systèmes d'information en santé à des fins de meilleure prise en charge du patient. L'ingénierie des modèles permet aux systèmes d'information le stockage et le partage d'information au sein d'un ou de plusieurs systèmes d'information hospitaliers, l'ingénierie de la connaissance permet la formalisation et l'intégration de la connaissance au sein du système d'information hospitalier (SIH) et l'intelligence artificielle permet de mettre en oeuvre des méthodes de raisonnement pour gérer la connaissance. L'informatique médicale est une science à part entière. Aux confluents des sciences de l'information et de la médecine, elle vise à proposer sa contribution pour la compréhension des mécanismes d'interprétation et de raisonnement médical, d'abstraction et d'élaboration des connaissances, de mémorisation et d'apprentissage. La science du traitement de l'information médicale [Degoulet 1998] touche aux fondements mêmes de la médecine. La figure 1.1 est une représentation graphique de l'interdisciplinarité de l'informatique médicale proposée dans [Baneyx 2007] dans laquelle nous avons ajouté les domaines des bases de données, d'ingénierie des modèles, de la fouille de données et d'aide à la décision, qui sont des domaines que nous aborderons au cours de ce mémoire. Cette représentation graphique vise à représenter l'interconnexion entre diverses communautés de recherche vis à vis du domaine de l'informatique médicale.

1.1.1 Pourquoi une science de l'information médicale ?

Abordons la question par une série d'interrogations. Comment stocker et réutiliser l'information médicale ? Comment modéliser la connaissance médicale pour qu'elle puisse être exploitée sans ambiguité ? Peut-on imaginer un système de traitement de l'information qui soit capable de gérer toutes les spécialités médicales

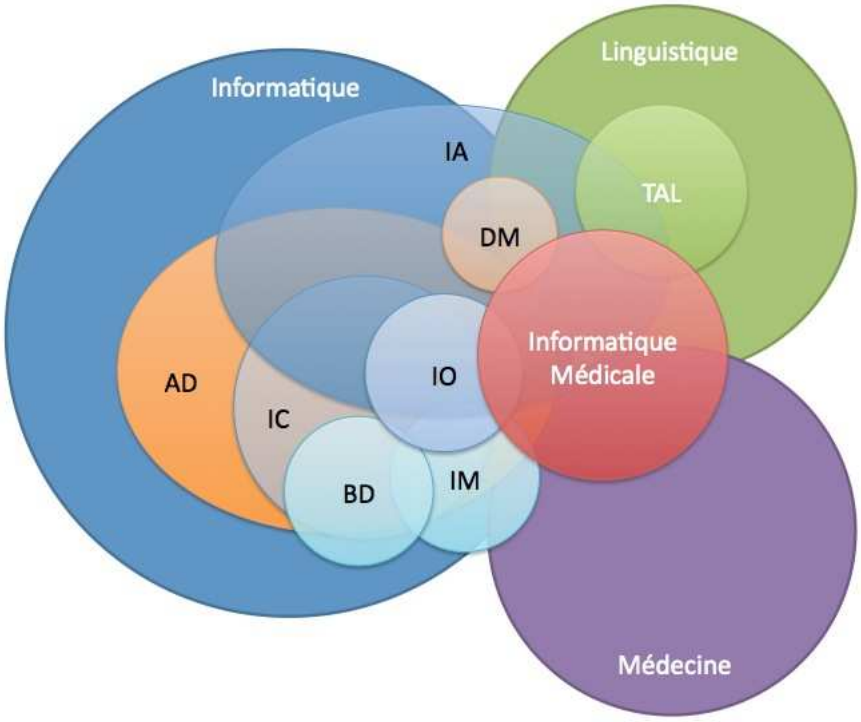


FIGURE 1.1 – Interdisciplinarité de l'informatique médicale. Version augmentée issue de [Baneyx 2007]. IM : Ingénierie des Modèles, BD : Bases de Données, DM : DataMining, IO : Ingénierie Ontologique, IC : Ingénierie des Connaissances, IA : Intelligence Artificielle, TAL : Traitement Automatique du Langage, AD : Aide à la Décision

(génétique, clinique, imagerie, moléculaire, biologique, etc.) ? Comment partager l'information médicale ? Quel est le processus qui permet de passer du signe au diagnostic puis à la décision médicale ? Peut-on définir une éthique du traitement de l'information médicale ? Existe-t-il une connaissance de référence toujours valide ? Comment intégrer des mécanismes biologiques et leurs interactions alors que nous n'en connaissons pas toute la nature ? Comment faire évoluer les connaissances des systèmes d'information ? Voilà quelques questions auxquelles les sciences de l'information médicale tentent de répondre afin de proposer et de mettre en oeuvre des solutions informatiques innovantes qui répondent aux exigences des personnels de la santé en termes d'utilisabilité, aux populations en termes de soin et de service rendu, et qui permettent aux chercheurs de découvrir de nouvelles connaissances. L'informatique "fondamentale" ne peut que s'intéresser partiellement aux problématiques de la médecine. En effet, les outils et méthodes issus des sciences de l'information, de l'intelligence artificielle ou bien du traitement automatique des langues ne peuvent

à eux-seuls, d'une part s'adapter au besoins du monde médical, et d'autre part créer les outils nécessaires à la gestion spécifique que demande l'information médicale. Le type d'information géré dans le monde de la santé est très complexe. De part son caractère hétérogène(textes, images, sons, ondes, radios, IRM, etc.), mais surtout de part sa volumétrie et sa complexité de représentation (interactions moléculaires, génétique, biologie, environnement, clinique, etc.). L'informatique médicale est donc une science appliquée à part entière qui, bien que récente, est aujourd'hui au coeur de la modernisation de la médecine [1].

1.1.2 L'évolution des architectures des systèmes d'information de santé

L'architecture des systèmes d'information en santé a suivi la même évolution que d'autres domaines (banque, industrie, etc...). D'une approche horizontale (architecture centralisée en étoile dans les années 70) où l'information est saisie une fois et est accessible depuis tous les postes, à une approche verticale (par département) où chaque département à sa propre application, l'architecture d'un SIH tend aujourd'hui à être distribuée ; les applications sont intégrées les unes aux autres par une approche de type grille orientée services. Par exemple, les hôpitaux universitaires de Genève ont été parmi les premiers à mettre en oeuvre une architecture orientée services [Lovis 2006] par fonction hospitalière et à quitter l'approche verticale. En France, nous pouvons aussi donner l'exemple de l'HEGP [2] qui présente un taux élevé d'informatisation et qui a une architecture SIH par service. Il est cependant à noter qu'il y a un manque d'informatisation globale des SIH. Historiquement, l'architecture centralisée permettait en aval, la mise en oeuvre d'outils d'analyse de données adhoc. Il était plus facile de récupérer l'information nécessaire à de l'analyse de données en un point unique. Aujourd'hui, pour faire de l'analyse de données, il est nécessaire d'intégrer des données reparties depuis différents services, et de mettre en oeuvre des solutions techniques qui permettent l'interrogation rapide de très grands volumes de données. Les entrepôts de données sont une réponse à ce besoin d'analyse, mais nous verrons dans ce mémoire de thèse que d'autres réponses sont possibles, suggérées par exemple par le développement du web sémantique.

1.1.3 L'informatisation actuelle

L'informatisation des établissements de soins est actuellement poussée par un aspect socio-économique. En effet, l'informatique peut permettre d'améliorer la

1. http://money.cnn.com/2009/01/12/technology/stimulus_health_care/
2. Hôpital Européen Georges Pompidou

communication inter-services, d'améliorer la gestion administrative d'un hôpital, de mieux gérer la pharmacie ou bien la radiologie. Cependant, pour arriver à ces objectifs, il faut saisir l'information médicale, la stocker pour pouvoir l'interroger et l'utiliser pour faire des analyses. Toute formalisation de pratiques à travers l'outil informatique a évidemment fait émerger de nouveaux besoins d'analyse des pratiques dans divers buts comme la rationalisation des coûts ou la diminution des risques. L'axe économique est un moteur de cette informatisation, mais il ne peut être le seul moteur. Alors qu'il est normal dans une communauté de s'interroger sur l'efficacité financière d'un système communautaire (en France : la sécurité sociale et le PMSI [3]), il est beaucoup plus difficile de juger de la qualité de décision d'un praticien ou d'un traitement seulement sur ces considérations. L'axe financier peut être un garde fou lorsque la médecine déploie des moyens colossaux dans certains cas d'acharnement thérapeutique là où le patient et les familles désirent mettre fin à cet acharnement. Mais il peut aussi être un danger quand on décide de politiques nationales en regardant principalement des chiffres. Outre la problématique de mesure financière d'une activité, la mise en oeuvre d'outils permettant de mesurer précisément l'activité d'un personnel de santé pose des questions déontologiques d'une part, et de respect des pratiques d'autre part. Il est, encore aujourd'hui, très difficile de modéliser dans des systèmes d'information l'exactitude et la richesse d'une pratique, quelle qu'elle soit, il est tout aussi impossible de capturer dans des systèmes informatiques toute la richesse de la médecine, des pratiques médicales, et des interactions médecin-patient de telle manière qu'on puisse mesurer avec justesse les subtilités de la pratique médicale.

1.2 Positionnement

Nous pensons cependant que le partage de l'information médicale devient aujourd'hui une nécessité dans un monde où la mobilité et les interactions humaines sont toujours plus importantes. Le partage d'information permet de construire des systèmes d'alertes à grande échelle, plus réactifs, qui ne nécessitent pas la mise en oeuvre de systèmes coûteux de recueil d'information qui sont bien souvent déjà obsolètes lorsque l'étude se termine. Être en mesure de partager des informations médicales de qualité, sans ambiguité, et de manière assez large et en temps réel, sous-entend qu'un certain nombre de verrous scientifiques soient levés. Intéressons nous au cycle de vie de l'information médicale pour mieux appréhender ces verrous :

– D'abord, le recueil des données de santé doit, autant que possible, se faire de

3. Programme de Médicalisation des Systèmes d'Information

manière codifiée et assistée tout en étant adapté à la prise en charge du patient par le personnel de santé.
- Ensuite, le stockage des données recueillies doit permettre l'évolution des données avec le temps, leur relecture ainsi que leur traçabilité.
- Les données recueillies doivent pouvoir être partagées sans ambiguité au sein même du système d'information et à l'extérieur de celui-ci.
- Elles doivent enfin pouvoir permettre la création de nouvelles connaissances sur le patient, un groupe ou une population de patients (réutilisation des données).

Il est difficile aujourd'hui de mettre en oeuvre ce cycle de vie avec tous les éléments nécessaires au partage de l'information. D'une part, il est trop fastidieux pour les professionnels de santé de coder toute l'information dans les SIH (il est difficile de sélectionner le code d'un signe clinique dans une liste de 20 000 items) pour des raisons de temps, mais aussi car les termes de ces listes n'ont pas nécessairement la sémantique qu'ils recherchent. D'autre part, il est difficile de stocker des données avec leur sémantique formelle. Il convient donc de trouver des méthodes et des outils permettant l'échange d'information médicale telle qu'elle a été entrée par les professionnels de santé dans les systèmes d'information.

1.3 Objectifs

C'est pourquoi dans le cadre de cette thèse, nous nous intéresserons particulièrement à la mise en oeuvre d'outils et de méthodes nécessaires au partage d'information dont le sens et la qualité sont connus et partagés. Nous aborderons la problématique du partage d'information suivant 3 axes : les modèles, la sémantique et la qualité. Nous proposerons une méthode d'évaluation de la qualité d'une source d'information pour l'interopérabilité. Nous proposerons ensuite une méthode de stockage de l'information et de la connaissance issue de l'ingénierie de l'informatique décisionnelle et du web sémantique. Nous aborderons ensuite la problématique de la mise en oeuvre d'une plateforme d'interopérabilité sémantique dans le cadre de l'échange d'information sur le domaine des maladies infectieuses. Cette plateforme devra permettre le partage(1) de connaissances médicales(2) de manière interprétable par la machine(3). Ces 3 dimensions relatives à l'interopérabilité sémantique seront mises en oeuvre dans le cadre du projet Européen DebugIT [4] dont l'objectif, tout comme cette thèse, sera de valider l'utilisation des technologies du web séman-

4. Detecting and Eliminating Bacteria UsinG Information Technology : Projet de surveillance et d'analyse de l'évolution de la résistance des bactéries aux antibiotiques en Europe.

tique (ou web de données [5]) dans le cadre du partage d'information biomédicale en fonction de la nature de l'information à partager.

1.4 Le contexte et l'expérimentation

L'évolution de la résistance aux antibiotiques en Europe devient alarmante [6]. Afin de réagir plus rapidement aux nouvelles résistances, une solution est d'avoir accès à l'information réelle plus rapidement. Pour avoir de l'information en temps réel afin de surveiller plus efficacement cette évolution, il est nécessaire d'utiliser des outils issus des technologies de l'information d'une nouvelle manière. Se "connecter" aux bases de données des hôpitaux informatisés ne suffit pas. Il faut pouvoir modéliser des langages différents, des vocabulaires différents, et même des concepts différents. Nous pensons que la communauté du web sémantique apporte des solutions dans ce cadre. DebugIT propose, entre autres, de mettre en place une plateforme de surveillance en se basant sur les technologies du web sémantique afin d'agréger l'information provenant de sources hétérogènes en temps réel de manière sécurisée grâce à l'Internet. Les travaux de cette thèse s'inscrivent dans ce contexte.

1.5 Enjeux scientifiques

Cette thèse est transversale à plusieurs domaines de recherches, fondamentaux ou appliqués. Elle n'est pas une contribution fondamentale aux domaines des bases de données, de l'ingénierie des connaissances ou de l'intelligence artificielle. Elle vise clairement à définir les méthodes nécessaires à l'échange d'information biomédicale de manière générale et plus particulièrement dans le domaine des maladies infectieuses. La plateforme d'interopérabilité, qui met en oeuvre les méthodes présentées dans ce travail est assez générique pour être utilisée dans d'autres domaines médicaux. Nous discuterons au cours de cette thèse du rôle des ontologies dans le domaine du partage d'information, leurs avantages et leurs limites. Nous aborderons le rôle et l'importance des modèles d'information et des vocabulaires nécessaires au partage d'information. Nous nous poserons aussi la question de la qualité des données pour l'interopérabilité, et particulièrement lorsque l'on veut utiliser des données dans des approches où données et sémantique sont couplées. Nous discuterons aussi

5. http://linkeddata.org/

6. Les Echos publie une enquête sur les antibiotiques, intitulée « chronique d'un désastre annoncé ». Laurence Bollack remarque en effet que « la résistance aux antibiotiques suscite une inquiétude grandissante dans le milieu médical. En cause, la surconsommation, les médicaments génériques. Et surtout l'abandon de la recherche... ».

de l'évolution nécessaire des modèles de données pour mieux prendre en compte la sémantique afin nous permettre d'espérer, un jour, de créer des outils d'intégration de données plus automatiques. Enfin, nous espérons que nous pourrons éclaircir le lecteur quand aux termes mêmes de l'interopérabilité et de la sémantique et des mythes qui semblent aujourd'hui être véhiculés dans diverses communautés de recherche. Il nous semble que le domaine de la santé est un excellent exemple afin de lever ces mythes.

Première partie

Partage de données et connaissances biomédicales

Chapitre 2

Modèles et Représentations

"Je pourrais comparer ma musique à une lumière blanche dans laquelle sont contenues toutes les lumières. Seul un prisme peut dissocier ces couleurs et les rendre visibles : ce prisme pourrait être l'esprit de l'auditeur." - Arvo Pärt, compositeur.

Sommaire

2.1 Introduction 14
2.2 Modèles de données 16
2.2.1 Un bref historique 17
2.2.2 Le modèle relationnel 18
2.2.3 Le modèle de données objet 19
2.2.4 Le modèle dimensionnel 21
2.2.5 Le modèle de données dimensionnel semi-structuré 22
2.2.6 Le modèle EAV 24
2.2.7 Le modèle dimensionnel EAV 25
2.2.8 Le modèle associatif 25
2.2.9 Le modèle RDF 28
2.2.10 Synthèse 31
2.3 Modèles de connaissance 33
2.3.1 Définitions 34
2.3.2 Terminologies et Classifications 35
2.3.3 Ontologies 37
2.3.3.1 Anatomie d'une ontologie 38
2.3.3.2 Natures d'ontologies 41
2.4 Modèles de qualité 43
2.4.1 La qualité de l'information 43
2.4.1.1 Les mesures de qualité 43
2.4.1.2 Les processus d'amélioration de la qualité 44
2.4.1.3 Exemples de projets qualité 45
2.4.2 Synthèse 46

2.5 Synthèse 46
2.6 Discussion 47

Des données et des connaissances sont stockées dans des systèmes d'information suivant des structurations bien précises. La manière dont est stockée l'information est souvent fonction de son usage, cependant, des structures de stockage ont gagné la faveur des communautés. Nous débutons donc notre état de l'art par l'étude des structures de stockage de l'information telles qu'elles sont proposées par les communautés des bases de données. Nous aborderons les différents modèles de données proposés au cours du temps, et nous observerons que l'évolution des structures de stockage va de pair avec l'usage de plus en plus courant de systèmes d'organisation de la connaissance. Dans un contexte de partage de l'information biomédicale, il est tout aussi essentiel d'aborder les différentes structures d'organisation de la connaissance puisqu'elles peuvent être différentes. Nous présenterons donc dans un deuxième temps différents modèles adaptés à l'organisation de la connaissance. Nous montrerons comment les structures de stockage influent sur la puissance expressive de l'information stockée, et donc sur la capacité de ces systèmes à interroger et partager l'information qu'ils détiennent. Enfin, bien que les modèles de données et de connaissance soient connus, nous pensons qu'il est nécessaire d'aborder la problématique de la qualité de l'information pour le partage. Nous présenterons donc ensuite succinctement ce que la communauté de la qualité de données peut apporter à notre problématique. Nous verrons dans la suite de cette thèse comment la qualité de l'information peut être vue comme une "variable d'ajustement" dans le cadre du partage de l'information.

2.1 Introduction

Pour illustrer le problème d'expressivité de l'information, parlons d'abord de mouton. Le mot "mouton" peut avoir le même sens que le mot Anglais "sheep" ; mais ils n'ont pas la même valeur. En effet, le mot Anglais pour désigner la viande de cet animal (que l'on cuisine) est "mutton" et non "sheep" alors qu'en Français, le même mot "mouton" couvre les deux valeurs. La figure 2.1 nous présente une manière de formaliser ce problème. On observera que cet exemple ne se réduit pas à un problème de multilinguisme simple. En effet, il est nécessaire tout d'abord de

structurer nos concepts en fonction de leurs concepts parents. Ici, nous différencions le concept de viande de mouton, de l'animal lui-même. La langue Française ne fait pas nécessairement la différentiation en terme de symbole entre les deux concepts du mouton. Outre l'aspect de l'historique linguistique qui pourrait expliquer cette différence, et qui, au demeurant, est très intéressant, nous remarquons que la langue anglaise est mieux adaptée pour modéliser la connaissance associée à notre domaine du mouton.

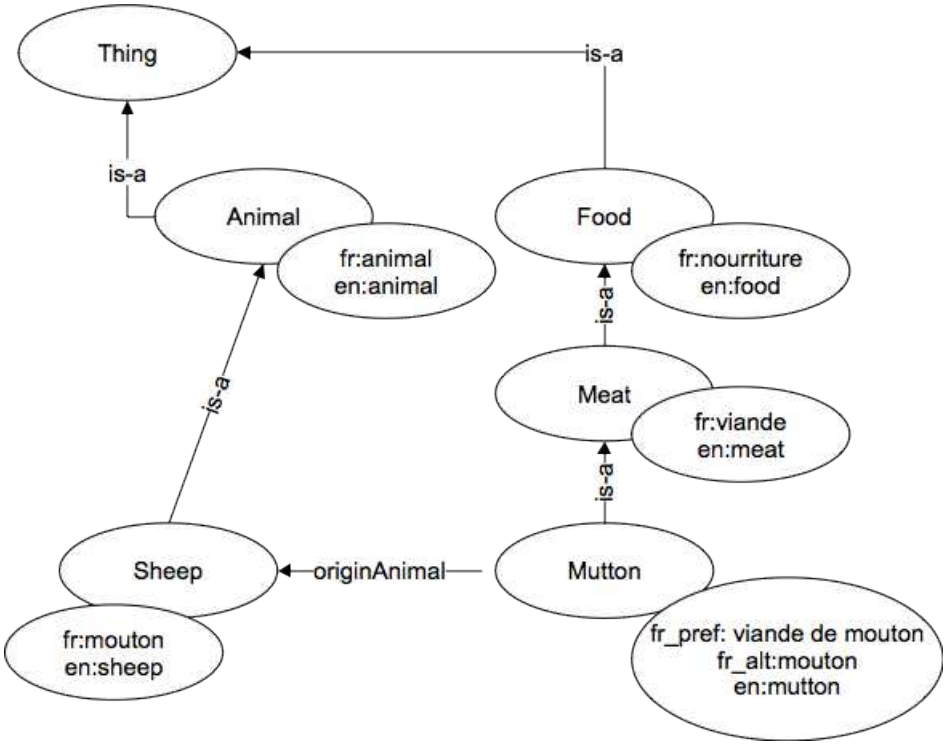


FIGURE 2.1 – *Le problème du mouton Anglais.* Dans cet exemple, nous modélisons les concepts Animal, Food, Meat, Mutton et Sheep comme des concepts abstraits du domaine ; on aurait pu utiliser des signes, ou des codes pour représenter les abstractions conceptuelles de notre domaine. Les labels sont représentés avec le préfixe de langage (fr, en) qui permet de gérer le multilinguisme. Il est aussi possible dans cette représentation ontologique de définir des termes préférés pour nommer un concept (fr_pref, fr_alt).

Le problème du mouton anglais met en avant plusieurs domaines connexes de recherche en sciences de l'information qui s'intéressent tous, de près ou de loin, au langage, à la modélisation des concepts du monde et aux relations qui existent entre ces deux domaines. Historiquement, l'ingénierie des modèles et des processus, puis plus récemment l'ingénierie de la connaissance, sont des disciplines qui tentent

de formaliser (pour la machine ou pour l'homme) la signification des termes que nous utilisons pour représenter les objets matériels ou immatériels et les relations ou interactions entre eux. Ces domaines nous ont apportés diverses méthodes et outils afin de nous aider à modéliser l'information et à la traiter. Nous présenterons d'abord les modèles d'information tels qu'ils ont été abordés par la communauté des bases de données. Nous observerons ensuite d'autres modèles d'organisation de la connaissance indispensables à l'organisation des données contenues dans les bases de données. Nous verrons que, bien que par le passé les modèles de stockage différaient, ils tendent aujourd'hui à chercher un mode commun de représentation.

2.2 Modèles de données

"Un modèle de données est un modèle qui décrit de façon abstraite comment sont représentées les données dans une organisation métier, un système d'information ou une base de données." - Wikipedia.
Les modèles d'information expriment une vue abstraite, dépendante de l'observateur, sur une réalité spécifique et avec un but précis. Chaque modèle conceptuel, logique ou physique peut avoir différents formalismes de représentation. Alors que les modèles conceptuels (tels que UML) donnent une assez grande liberté d'expression, il est communément admis que les modèles logiques de données (puis physiques) posent un cadre limitant et réduisent l'expressivité du domaine étudié. Très souvent, les problèmes de performance d'accès aux données en sont la cause.

L'histoire des modèles d'information a très vite confronté deux grandes familles de modèles. D'une part les modèles relationnels et d'autre part les graphes conceptuels. Les modèles relationnels, adaptés à la gestion de grandes masses de données, tentent peu à peu de s'adapter aux nouveaux usages que l'Internet suggère, à savoir l'utilisation d'une masse de données réparties dans un grand graphe global, mais sans grand succès. La structure même de l'Internet, et la pauvreté sémantique du modèle relationnel nous pousse à penser que ce modèle n'a pas les propriétés nécessaires pour aborder les problématiques de demain. En parallèle, nous assistons aujourd'hui à l'avènement de la conceptualisation de l'information en graphes grâce notamment à RDF [1]. Nous aborderons dans ce chapitre les notions de modèles de données issus de l'ingénierie des modèles. En premier lieu pour le stockage de données, puis, pour l'analyse avec les modèles dimensionnels, enfin, nous présenterons les notions de modèles en réseaux et en arbres utilisés historiquement pour la gestion de la connaissance. Nous remarquerons l'évolution de ces modèles vers des modèles

1. Resource Description Framework. RDF est un modèle de stockage de l'information

où connaissance et données peuvent cohabiter. Nous catégoriserons ces modèles en fonction de leurs usages et de leurs performances.

2.2.1 Un bref historique

La fin des disques magnétiques pour stocker de l'information a contribué à créer de nouveaux modèles de stockage non sérialisés. La recherche a, dans les années 70, aboutit à 3 types de modèle d'information : le modèle hiérarchique, le modèle en réseau et le modèle relationnel. Ce dernier sera mis en avant par IBM pour la gestion des données dans des bases de données. Plus tard, le modèle objet-relationnel et le modèle de données semi-structuré seront développés. Ces modèles logiques [2] permettront de stocker de grandes quantités d'information sur des disques magnétiques (disques durs) et la recherche et l'industrie mettront en oeuvre des outils permettant de vérifier l'intégrité des données ainsi que la recherche rapide d'information grâce aux moteurs de bases de données. Le modèle logique relationnel basé sur des modèles conceptuels comme Merise est très adapté pour le stockage de l'information et pour sa mise à jour. Pendant la première vague de numérisation de l'information, il sera le modèle principalement utilisé pour les usages communs des entreprises facilement numérisables comme par exemple la gestion de la comptabilité ou de la paie. Il a rapidement été possible de stocker de grandes quantités d'information, de les retrouver, de les modifier avec de bonnes performances. L'histoire montrera que ces systèmes étaient en fin de compte ce qu'on appellerait aujourd'hui dans le domaine biomédical : des systèmes de capture d'information, reliés à une interface homme-machine, où l'homme pouvait passer des transactions d'écriture, de lecture ou de modification de ses données numériques. Une fois que ces masses d'information se sont constituées, nous nous sommes heurtés à la problématique de la recherche d'information, de performance et d'intégration multi-sources pour l'analyse de ces données. C'est dans ce contexte qu'un modèle conceptuel et logique de données sera introduit dans les années 90 : le modèle dimensionnel. Il sera d'abord adapté aux bases de données relationnelles, puis sera adapté aux bases semi-structurées, aux bases objet et récemment aux bases de données RDF. Ce modèle permettra d'offrir une structuration de l'information adaptée à l'analyse, où, nous le verrons, la notion de contexte et de sens apparaîtra. Au cours de ces évolutions, nous remarquerons quelques tentatives d'adaptation du modèle relationnel dans des variantes où la structure du modèle ne nécessitera plus de modification en cas de changement de périmètre conceptuel, mais cela se fera toujours au prix des performances

2. Un modèle logique est une représentation d'un modèle conceptuel de données (par exemple le modèle Entité-Relation) compréhensible par la machine et le moteur de base de données

d'interrogation.

2.2.2 Le modèle relationnel

Ted Codd, dans [Codd 1970, Codd 1979], propose un modèle relationnel de données pour apporter plus d'indépendance aux éléments de données en opposition aux modèles en réseau et en graphes. Il propose de représenter les données en les structurant dans des tables contenant des colonnes et des tuples ; le nom de la table définissant le domaine de l'information stockée. Les relations entre les tables, effectuée par un mécanisme de clé primaire - clé étrangère, permettent de mettre en relation un ou plusieurs éléments (n-aires). Une relation au sens de Codd, ne peut cependant pas représenter de relation sémantique entre deux concepts puisqu'elle n'est pas stockée physiquement dans la base de données. Elle représente par contre un concept mathématique comme Codd l'a définit :

Definition 1. *Sachant les ensembles E_1, E_2, ... , E_n, R est une relation portant sur n ensembles si c'est un ensemble de n-uplet, le premier élément appartenant à E_1, le deuxième à E_2, etc. Un tableau qui représente une relation n-aire R a les propriétés suivantes :*

- *chaque ligne représente un n-uplet de R,*
- *chaque ligne est distincte,*
- *l'ordre des ligne est signifiante, il correspond à l'ordre des ensembles E_1, ... , E_n des domaines dans lesquels R est défini,*
- *le sens de chaque colonne est partiellement apporté par le label de celle-ci*

Par exemple, la table 2.1 illustre une relation de degré 4 nommée *provision* reflétant l'acheminement de produits depuis des fournisseurs dans un projet spécifique et en quantité donnée.

provision	(fournisseur	produit	projet	quantité)
	1	2	5	17
	1	3	5	23
	2	3	7	9
	2	7	5	4

Table 2.1 – Exemple de représentation tabulaire d'une relation de degré 4

La relation sémantique entre les différents éléments d'un domaine ne sera pas matérialisée dans le modèle relationnel. La relation proposée est une relation qui restera structurelle et qui est généralement interprétée par l'homme pour en comprendre

son sens. Dans un langage permettant la mise en oeuvre de modèles conceptuels de plus haut niveau (par exemple UML), il est possible de caractériser plus finement la relation entre deux éléments, mais plus à un niveau informatif que sémantique. Ce type de relation informative ne permettra pas la mise en oeuvre de processus d'inférence logique [3] puisque la relation n'est pas explicitement stockée.

L'exemple 2.2 montre la modélisation relationnelle concernant le stockage d'information relative à des patients dans des hôpitaux qui ont été diagnostiqués. (Nous utiliserons cet exemple simple dans la suite de ce chapitre pour illustrer les différences entre les modélisations.) La cardinalité est représentée par le sens des flèches. Il faut lire par exemple : *Un patient est dans 1 hôpital, et un hôpital a plusieurs patients.*

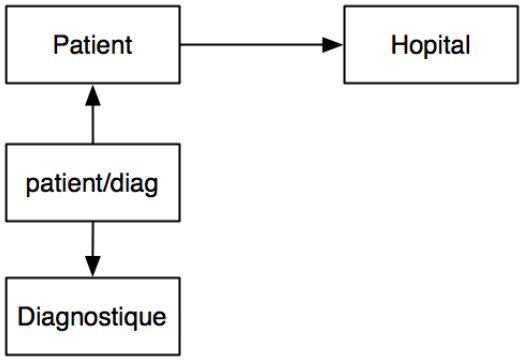


FIGURE 2.2 – Le modèle relationnel permet la mise en oeuvre de contraintes sur les relations. Il n'est par contre pas possible de stocker la sémantique des relations entre les éléments.

2.2.3 Le modèle de données objet

Le modèle de données objet, inspiré des méthodes de programmation objet, a été introduit dans les années 80. Il vise à se rapprocher des outils de modélisation conceptuelle tels que UML et propose, via le paradigme objet, une gestion de l'information plus proche des concepts (à priori) du monde réel. Ce modèle de données n'a pas été grandement adopté par l'industrie contrairement au modèle relationnel qui aujourd'hui encore est le modèle le plus utilisé dans le monde. Une des raisons peut être le manque de formalisme du modèle objet de données à ses débuts, ou alors sa complexité de prise en main au regard du service rendu (sachant que les méthodes

3. L'inférence est une méthode de déduction de connaissances pouvant être mis en oeuvre sur des réseaux de connaissances formellement définis

d'accès aux données sont souvent en dehors du SGBD). Dittrich propose cependant que le modèle objet de données adopte les propriétés suivantes : *persistance, gestion de stockage secondaire, récupération et existence d'un système de gestion de requêtes* ; mais aussi : *la gestion d'objets complexes, d'identification des objets, d'encapsulation, de types ou de classes, d'héritage, d'extensibilité, ...* [Dittrich 1986].

Les objets sont des représentations du monde réel, au même titre que des concepts dans une ontologie, à la différence que chaque objet encapsule ses données (propriétés) et ses comportements (méthodes). Chaque objet est indépendant. Les liens entre les objets sont matérialisés par des relations spécifiques comme des associations ou un mécanisme d'héritage. Ils peuvent stocker la relation sémantique comme propriété, cependant les relations ne sont pas stockées de manière indépendantes de l'objet, les rendant plus informatives qu'utiles pour de l'inférence par exemple. L'exemple 2.3 reprend l'exemple précédemment vu en modélisation objet formalisé en UML. Un attribut de classe est par exemple le Nom ou le Prénom d'un patient. Une opération représente par exemple l'ajout ou la modification d'un patient à l'ensemble des patients.

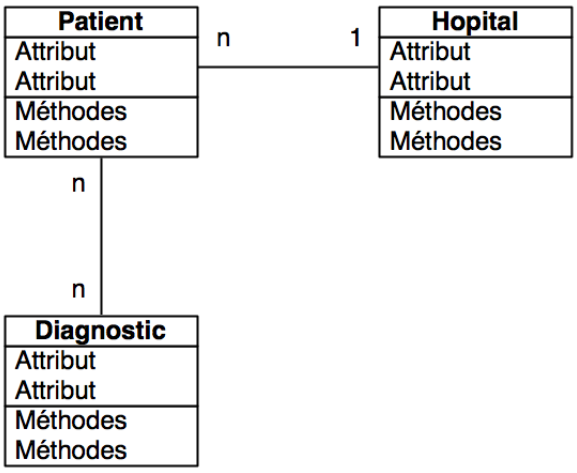


FIGURE 2.3 – Le modèle objet permet la mise en oeuvre de contraintes sur les relations ainsi que la mise en oeuvre de méthodes associées à chaque élément. Il n'est par contre pas possible de stocker le sens des relations entre les éléments directement. Une technique est cependant implicitement possible en mettant en oeuvre des méthodes d'accès aux données adéquates.

Les modèles relationnels-objets ont été introduits dans les années 90. Ils tentent d'apporter le meilleur des deux mondes, à savoir la capacité d'interrogation et la gestion des données complexes. Stonebraker propose une classification des systèmes

de gestion de bases de données [Stonebraker 1996]. Nous proposerons une vue élargie de la classification de Stonebraker dans la figure 2.12 en conclusion de cette section de chapitre.

2.2.4 Le modèle dimensionnel

Le modèle dimensionnel fût introduit par Codd qui définit déjà les concepts de relation et de dimension. Le modèle en étoile (qui représente un cas particulier du modèle dimensionnel) fût défini par [Kimball 1995, Kimball 2002], puis fût formalisé par Inmon dans le cadre d'utilisation d'entrepôts de données dans un système d'information dans [Inmon 1995]. Ce modèle d'abord logique deviendra à l'usage un modèle utilisé dans les phases d'analyse conceptuelle. Ce modèle est le modèle de prédilection pour interroger de très grandes quantités de données avec un temps de réponse très rapide. Il est cependant peu adapté pour la mise à jour et la gestion quotidienne de traitements (transactionnels). Le modèle dimensionnel est aujourd'hui le modèle le mieux adapté pour l'analyse de données car il permet une certaine contextualisation des éléments d'un domaine. La structure de celui-ci peut être en étoile : *(1)table de faits - (n)dimensions*, en flocon de neige : *(1)table de faits - (n)hiérarchies de dimensions* ou en constellation : *(n)tables de faits - (n) hiérarchies de dimensions*. La table de *faits* représente l'élément que nous voulons mesurer (des ventes, des prescriptions, des diagnostics). Les dimensions représentent des éléments de contexte de mesure (un lieu, un vendeur, un patient, un médecin, une date). Des tables de faits peuvent partager une ou plusieurs dimensions. Le modèle dimensionnel est défini de la manière suivante dans [Boussaid 2006] :

Definition 2. *Soit* $D = \{D_s, 1 \leq s \leq r\}$ *un ensemble de* r *tables de dimension indépendantes. Chaque table* D_s *a une clé primaire* $D_s.PK$*.* F *est une table de faits comprenant* d *clés* K *multiples. Un "schéma en étoile" est défini par le couple* (F, D) *qui satisfait les conditions suivantes :*

- $\forall t \in \{1, ..., d\}$*, il existe exactement un* $s \in \{1, ..., r\}$ *tel que* $F.K_t = D_s.PK$ *;*
- $\forall s \in \{1, ..., r\}$*, il existe exactement un* $t \in \{1, ..., d\}$ *tel que* $F.K = D.PK$*.*

Lorsque le modèle dimensionnel (du point de vue conceptuel) est mis en oeuvre dans une base de données relationnelle (au niveau logique), il s'appuiera sur les relations structurelles de celui-ci (pauvres en sémantique) pour relier les éléments entre eux. Bien que la table de faits (dans l'exemple *estDiagnostiqué*) permette de rendre compte du contexte, la sémantique de celle-ci n'est pas formelle. Ici encore, seule une interprétation humaine pourra déduire les relations entre les tables de faits et les dimensions. Les relations entres les éléments dans les hiérarchies de dimension

sont de type 'is-a' [4]. L'exemple 2.4 montre la modélisation du patient grâce à la modélisation dimensionnelle en étoile. La table de faits *estDiagnostiqué* aura autant de tuples que de diagnostics posés. Chaque acte de diagnostic sera contextualisé grâce au patient, au diagnostic et à l'hôpital.

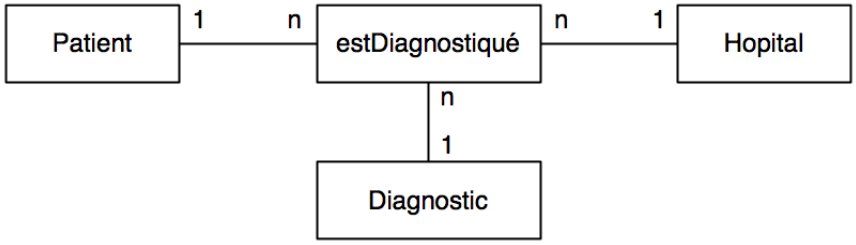


FIGURE 2.4 – Le modèle dimensionnel est fortement contextuel.La table *estDiagnostiqu* est renseignée de tuples qui ne peuvent exister que si un patient, un diagnostic et un hôpital sont renseignés.

Le modèle dimensionnel présente une approche intéressante pour observer des données dans un contexte donné de manière très performante. En effet, il s'appuie sur l'utilisation d'un mécanisme de filtrage des données basé sur des indexes sur la table de faits et sur les tables de dimensions. Les dimensions étant normalisées donc peu volumineuses, il est plus aisé pour le SGBD [5] de sélectionner un sous ensemble de la table de faits en utilisant le mécanisme de jointure.
Le modèle dimensionnel propose aussi une méthode de généralisation-spécialisation adapté aux SGBD relationnels. Les tables de dimension peuvent être découpées en hiérarchies. Par exemple, la dimension *temps* présente une hiérarchie claire (voir figure 2.5).

2.2.5 Le modèle de données dimensionnel semi-structuré

L'expansion de l'utilisation du formalisme XML [6] comme format d'échange a poussé la communauté des bases de données à proposer des moteurs de SGBD XML natifs parmi lesquels eXist [7], BaseX [8] ou encore Berkeley DB XML qui a été rachetée par Oracle. La structure en arbre d'XML peut être plus adaptée à l'exploitation de certains types de données. La structure XML est cependant pauvre sémantiquement

4. Une relation hiérarichique de type is-a permet de mettre en oeuvre une relation de type généralisation-spécialisation entre deux concepts d'un même domaine, voir Chapitre 2.4.3.1
5. Système de Gestion de Bases de Données
6. eXtended Markup Language
7. http://exist.sourceforge.net/
8. http://www.inf.uni-konstanz.de/dbis/basex/

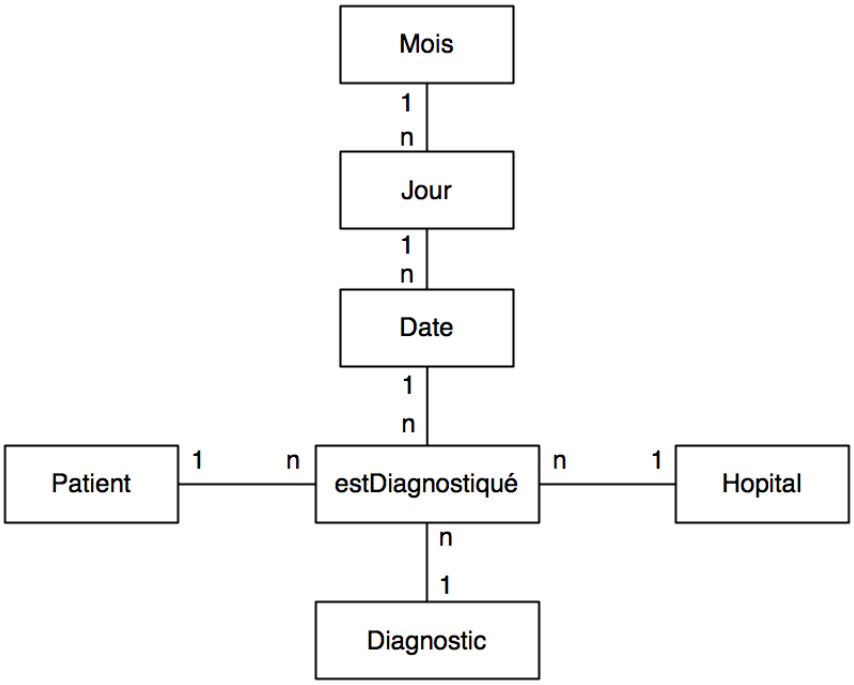


FIGURE 2.5 – Le découpage de la dimension temps dans une hiérarchie spécialisation-généralisation permet de regrouper les mesures de la table de faits (par exemple : un diagnostic positif) suivant l'axe temps en utilisant ici aussi l'optimisation du SGBD

pour exprimer la relation entre deux noeuds de l'arbre XML. L'utilisation d'XML pour modéliser et stocker des données dans le cadre d'entrepôts de données est abordé dans [Hümmer 2003, Pokorny 2001, Nassis 2004, Baril 2003]. Nous avons proposé dans [Boussaid 2006] une définition d'un schéma en étoile XML :

Definition 3. *Soit (F, D) un schéma en étoile où F est une table de faits ayant comme mesures m tel que $\{F.M_q, 1 \leq q \leq m\}$ et $D = \{D_s, 1 \leq s \leq r\}$ est un ensemble de tables r de dimensions indépendantes où chaque D_s contient un ensemble de n_s attributs $\{D_s.A_i, 1 \leq i \leq n_s\}$. Le schéma en étoile XML de (F, D) est un schéma XML tel que :*

- *F définit l'élément XML racine dans le schéma XML ;*
- *$\forall q \in \{1, ..., m\}$, $F.M_q$ définit un attribut XML inclus dans l'élément racine F ;*
- *$\forall s \in \{1, ..., r\}$, Ds définit autant de sous-éléments XML nécessaires liés à la table de faits F ;*
- *$\forall s \in \{1, ..., r\}$ and $\forall i \in \{1, ..., ns\}$, $D_s.A_i$ définit un élément XML attribut inclus dans l'élément XML D_s.*

Le modèle de données semi-structuré est particulièrement adapté à la modélisation de documents multimédia. Une modélisation semi-structurée de notre exemple précédent serait la suivante :

```
<?xml version="1.0" encoding="ISO-8859-1"?>
  <estDiagnostique>
    <Patient> </Patient>
    <Hopital> </Hopital>
    <Diagnostic> </Diagnostic>
    <Date date="">
      <Jour> </Jour>
      <Mois> </Mois>
    </Date>
  </estDiagnostique>
```

Le modèle de données semi-structuré que nous avons adapté pour des usages d'analyse au modèle dimensionnel permet la mise en oeuvre d'entrepôts de données XML pour l'analyse.

2.2.6 Le modèle EAV

Afin de pallier les problématiques liées à la complexité de la représentation des données biomédicales dans les systèmes de gestion de bases de données, le modèle EAV (entity-attribute-value) est proposé dans [Dinu 2007]. Dans un modèle EAV, la structure du modèle en base est indépendante de la structure des données issues d'un modèle d'information. Un enregistrement de type EAV est défini par :

- l'entité : l'élément d'information (l'objet),
- l'attribut : une clé vers une table des attributs que peut avoir l'entité,
- la valeur : la valeur de l'attribut.

Le modèle physique d'un EAV est généralement composé de quelques tables fixes mais peut être modélisé en une seule table contenant des triplets comme dans l'exemple suivant :

```
( <patient 1, 01/10/2010 09:30 AM>, <Diagnostic>, "Bronchite")
( <patient 1, 01/10/2010 09:30 AM>, <Hôpital>, "HEGP")
( <patient 1, 01/10/2010 09:30 AM>, <Nom>, "Baucuse")
```

Dans le cas de plusieurs tables, ces triplets sont des clés étrangères liées à des tables 'entity', 'attribut' et éventuellement 'value'.

Cette modélisation s'approche des modélisations associatives et en triplets (par exemple RDF, voir section suivante). La problématique générale de ce type de modélisation est le caractère implicite du stockage lorsqu'on interroge le SGBD avec le langage de requête actuel (SQL). En effet, il faut lire le contenu des tables pour connaître les éléments structurels qui permettent de lier l'information et donc de les représenter dans la requête. Aussi, ce modèle ne permet pas de décrire de manière formelle les relations entre les entités. L'autre problème des représentations EAV est la performance. En effet, les SGBD actuels ne mettent pas en oeuvre des index adaptés à ce type de représentation.

2.2.7 Le modèle dimensionnel EAV

Dernièrement, dans le domaine biomédical, un projet de recherche initié à Harvard à proposé une solution de stockage de données (i2b2 [9]) se basant sur un modèle dimensionnel simple mais ayant les propriétés d'un modèle EAV ou d'un modèle pragmatique [Ruelland 2003], à savoir la simplicité de structuration du modèle physique de la base de données, et la gestion du modèle logique dans des tables de métadonnées. La particularité du modèle dimensionnel-EAV d'i2b2 est qu'il apporte de la performance à celui-ci, mais ne rend pas plus explicite le modèle d'un point de vue de la structure. La table de fait du modèle est ici l'observation médicale au sens général, que l'on spécifie en annotant l'observation d'un terme que l'on trouvera dans une classification qui est stockée dans la dimension *concept_lookup* dans une hiérarchie. Il reste que les relations entre les observations ne sont pas modélisables et il est nécessaire de créer des clés de groupage des observations dans la table de faits.

L'approche i2b2 reste intéressante mais il nous apparaît que les relations sont tout d'abord toujours implicites, et même, dans ce cas, difficiles à modéliser.

2.2.8 Le modèle associatif

Suivant le courant des modèles de type réseaux sémantiques, Abrial [Abrial 1974] puis Bracchi [Bracchi 1976] présentent les associations binaires logiques dans le domaine de la modélisation de données afin de les enrichir sémantiquement. Ce modèle binaire de données est la base du modèle associatif ; il en diffère car le modèle associatif permet d'avoir des relations sur des relations [Williams 2001]. Le modèle associatif a été mis en oeuvre dans deux moteurs de bases de données associatives :

9. Informatics for Integrating Biology and the Bedside, https ://www.i2b2.org. Site accédé le 27/03/2008.

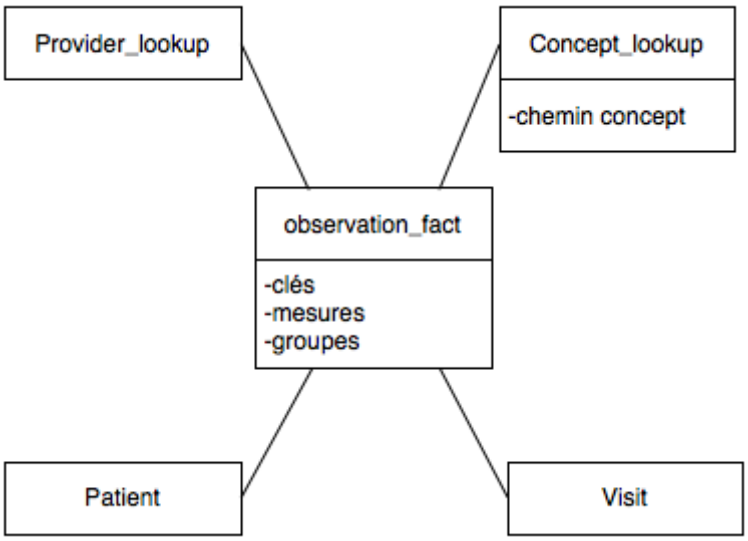


FIGURE 2.6 – Le modèle dimensionnel d'i2b2 s'apparente à une structure verticale de type EAV.

Sentences (Lazy Software©) et Relevance (Associative Solutions©). Plus récemment, QlikTech©a mis en oeuvre cette modélisation dans son outil d'analyse de données QlikView©. Enfin, le modèle associatif a été utilisé dans un projet de recherche en bioinformatique [Hanke 1999] afin d'aider à la visualisation de relations entre séquences génomiques sur une carte topographique en deux dimensions.

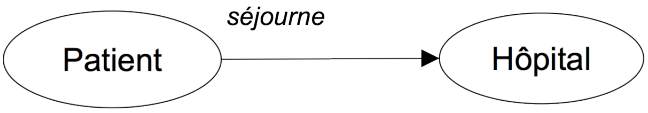


FIGURE 2.7 – Une association simple

Le modèle associatif est composé de deux types de structures : les entités et les verbes (ou relations ou associations sémantiques). Dans la figure 2.7, nous observons deux entités (patient et hôpital) ainsi qu'une association entre ces entités (séjourne). Dans cette association, le *Patient* est une entité puisqu'il a une existence indépendante, tout comme *Hôpital*. *séjourne* est par contre une association puisqu'elle n'exprime pas une entité indépendante. Pour étendre ce modèle, avec par exemple la notion de *diagnostic*, nous créons une association *est_ diagnostiqué* sur l'association définie précédemment (*séjourne*) comme dans la figure 2.8.

Il est aussi possible d'exprimer ce type d'association dans un langage naturel de la manière suivante :

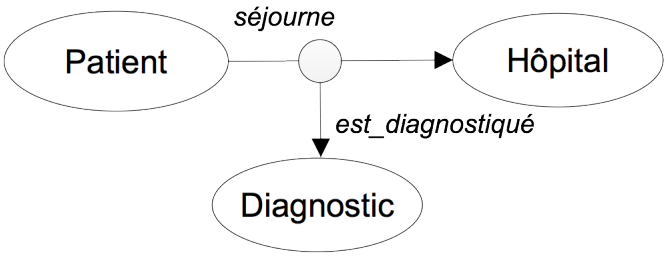


FIGURE 2.8 – Une association sur une association

Patient *séjourne* **Hôpital**
... *est_ diagnostiqué* **Diagnostic**

Une fois le modèle défini, il peut être instancié :
Paul Dupont *séjourne* **HEGP**
... *est_ diagnostiqué* **Infection Urinaire**

De plus, il est possible de définir des types permettant la spécialisation ou la généralisation d'entités. Par exemple : **Escerichia Coli** is a subtype of **Bacteria**. Des mécanismes d'héritage sont également possibles.

Dans la suite des définitions proposées précédemment, nous proposons la définition suivante du modèle associatif :

Definition 4. *Soit* $E = \{E_r, 1 \geq r \geq u\}$ *un ensemble d'entités indépendantes* u*, soit* $V = \{V_s, 1 \geq s \geq v\}$ *un ensemble de verbes indépendants* v*, soit* $A = \{A_t, 1 \geq t \geq w\}$ *un ensemble d'associations indépendantes* w *où* A_t *est défini dans une forme simple par :*

$A_t = E_r \circ V_s \circ E_{r'}$ *où* r' *peut être équivalent à* r *dans certains cas, donc :* $A_t = E_r \circ V_s \circ E_r$*,*

et où A_t *peut être encapsulé dans (une association sur une association) :*

$A_{t'} = A_t \circ V_{s'} \circ E_{r'}$
ou bien $A_{t'} = (E_r \circ V_s \circ E_r) \circ V_{s'} \circ E_{r'}$*.*

L'opérateur $\circ$ *matérialise le lien sémantique entre une entité et un verbe. Il exprime le contexte d'une entité.*

Le modèle associatif représente une des premières tentatives de matérialisation des relations entre les objets (concepts ou entités) modélisés au niveau du modèle logique. Il a été présenté juste après RDF.

2.2.9 Le modèle RDF

Resource Description Framework (RDF) est, comme le modèle associatif, un modèle de graphe orienté qui est destiné à l'origine, à décrire de manière formelle les ressources du Web ainsi que leurs métadonnées. Il a été développé par le W3C et publié en 1999 puis révisé en 2004 dans des spécifications [10]. Le modèle est devenu un modèle d'échange de données (métadonnées ou données) à part entière grâce à sa simplicité, à priori, de mise en oeuvre, mais aussi grâce à la puissance de représentation qu'il véhicule grâce aux graphes. RDF, comme le modèle associatif, permet d'exprimer explicitement les relations entre 2 ressources. RDF est sérialisable avec plusieurs syntaxes : XML, n3, OWL, etc. Le modèle RDF peut aussi être écrit à l'aide de triplets dans la forme *sujet - prédicat - objet* où chaque déclaration peut être une URI [11], une variable ou une valeur. Un sujet, un prédicat ou un objet peuvent être des URI, des littéraux ou des noeuds vides (ou noeuds virtuels). Par exemple :

```
<http://www.example.org/index.html>
<http://purl.org/dc/elements/1.1/creator>
<http://www.example.org/staffid/85740> .
```

Ce **triplet** [12] décrit une ressource (page web : http ://www.example.org/index.html) qui a été créée (http ://purl.org/dc/elements/1.1/creator)par une personne ayant pour identifiant 85740 (http ://www.example.org/staffid/85740). Le graphe obtenu sera orienté *sujet* vers *objet*. L'exemple suivant montre un graphe orienté décrivant des information sur moi-même où nous utilisons la resource *contact* de "http ://www.w3.org/2000/10/swap/pim/" afin de récupérer les propriétés (Person, fullName, mailbox, personalTitle) d'une *Person*, Rémy Choquet.

```
 <?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
             xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
  <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
    <contact:fullName>Rémy Choquet</contact:fullName>
    <contact:mailbox rdf:resource="mailto:remy.choquet@gmail.com"/>
    <contact:personalTitle>Dr.</contact:personalTitle>
```

10. http://www.w3.org/RDF/

11. Unified Resource Identifier : l'URI est un processus d'identification d'une ressource, un "identifiant web"

12. Un triplet est l'unité la plus petite pouvant décrire 2 noeuds d'un graphe et la relation entre ces 2 noeuds.

```
  </contact:Person>
</rdf:RDF>
```

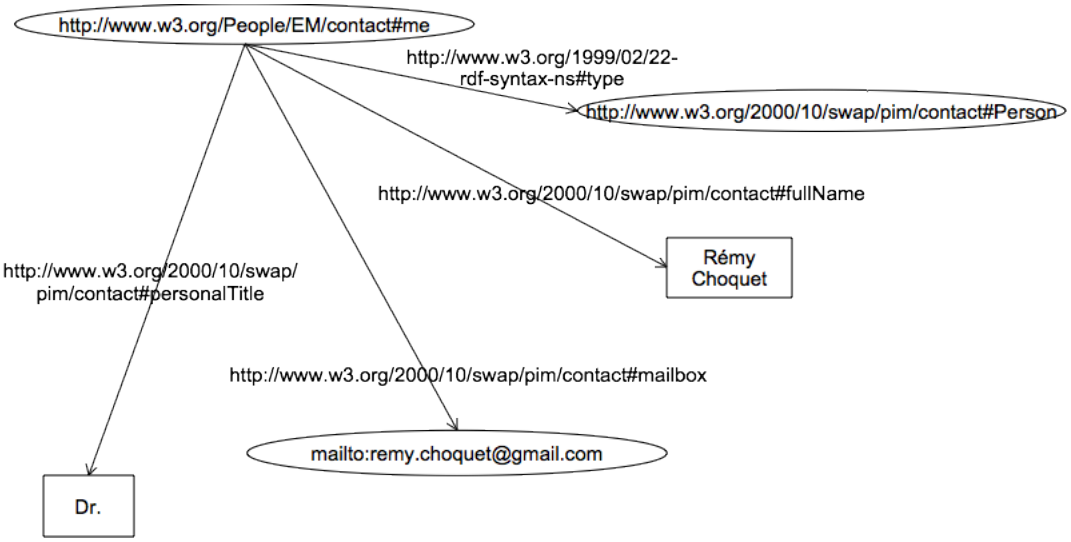


FIGURE 2.9 – Un graphe décrivant Rémy Choquet

Un mécanisme existe pour raccourcir les URI : les *QName*. Les QNames sont des préfixes permettant par exemple d'exprimer que `ex:` est un QName préfixant l'espace de noms URI de : `http://www.example.org/`. Le triplet suivant :

```
ex:index.html  dc:creator          exstaff:85740 .
ex:index.html  exterms:creation-date  "August 16, 1999" .
ex:index.html  dc:language            "enUS" .
```

L'avantage d'utiliser des URI pour référencer les éléments est que cela enlève toute ambiguïté quand au sujet ou à l'objet physique ou non décrit dans un triplet. Par exemple, ici, le créateur ayant pour identifiant http ://www.example.org/staffid/85740 ne peut être qu'une seule personne, même en cas d'homonymie. Il en va de même pour des concepts qui n'instancient pas un objet du monde réel, par exemple le langage "enUS" est l'anglais américain au sens où il est défini dans le Dublin Core [13]. Concernant la définition formelle de la sémantique des éléments, RDFS (RDF Schema) ou bien OWL(Ontology Web Langage) sont plus adaptés.
La modélisation de structures plus complexes (n-aires), qui est la plus courante dans

13. `http://dublincore.org/`

la réalité, peut se faire de plusieurs manières en RDF. Par exemple, pour décrire l'adresse postale d'une personne, il est nécessaire de la séparer en plusieurs éléments distincts tels que la rue, le numéro, la ville, etc. Ces éléments peuvent être associés à un élément adresse ou bien un noeud vide. L'usage d'un élément adresse de type identifiant universel (URI) est la modélisation la plus formelle possible, cependant, il n'est pas toujours utile d'aller jusque là ; on utilisera dans ce cas un noeud vide. Les figures 2.10 et 2.11 présentent deux manières différentes de modéliser des relations n-aires.

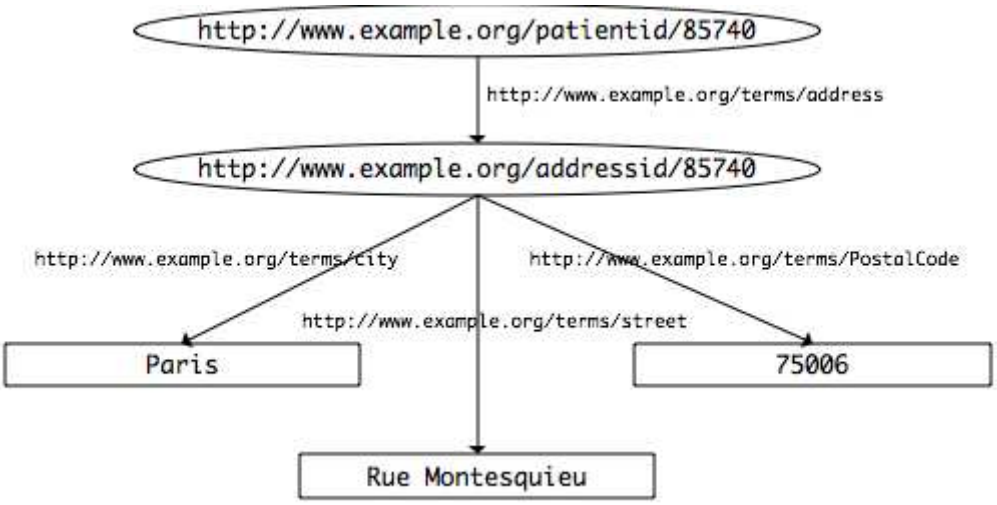


FIGURE 2.10 – Une relation n-aire avec un identifiant universel pour l'adresse

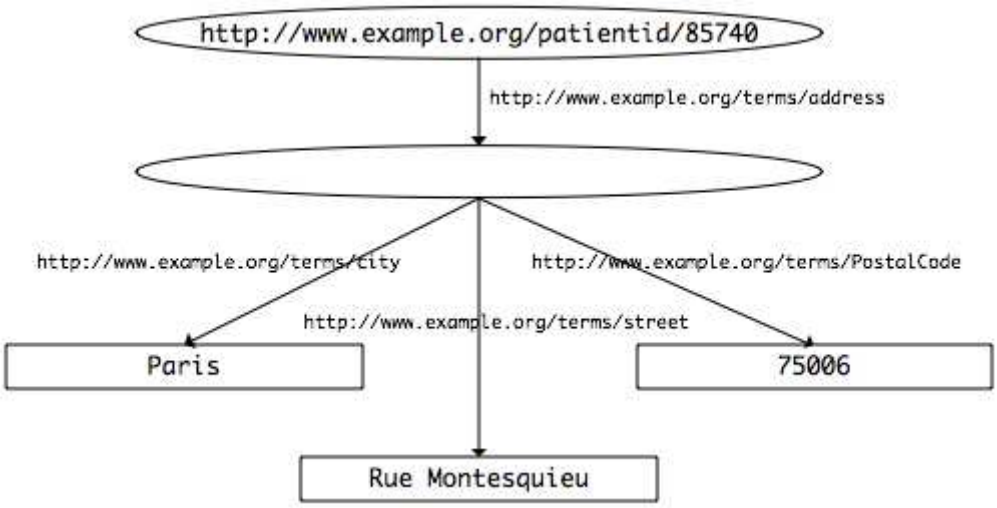


FIGURE 2.11 – Une relation n-aire avec un noeud vide.

Les triplets suivants représentent le graphe précédent avec l'élément "address" universel :

```
expatient:85740        exterms:address        exaddressid:85740 .
exaddressid:85740  exterms:city                 "Paris" .
exaddressid:85740  exterms:street             "Rue Montesquieu" .
exaddressid:85740  exterms:PostalCode  "75006" .
```

Les triplets suivants représentent le graphe avec le noeud blanc (ou virtuel) :

```
expatient:85740        exterms:address         _:pauladdress .
_:pauladress             exterms:city                 "Paris" .
_:pauladress             exterms:street             "Rue Montesquieu" .
_:pauladress             exterms:PostalCode  "75006" .
```

Dans le cas d'utilisation d'un noeud vide, ces noeuds ne sont pas considérés comme étant partie intégrante du graphe RDF résultant. Le mécanisme d'identification des noeuds vides (pauladdress) ne servant qu'à identifier un noeud blanc d'un autre.

2.2.10 Synthèse

Nous distinguons historiquement trois structures de stockages (modèles d'informations) différentes pour stocker l'information. Bien que celles-ci partagent des propriétés communes, ces structures ont des points forts et des points faibles. Le tableau 2.12 synthétise les différences actuelles entre les structures relationnelles, en arbre ou en graphe, en fonction de leur capacités de stockage, de recherche et de partage d'information. Nous avons mis en avant pour chacune de ces capacités quelques propriétés, nous avons qualifier chacune des structures en fonction de celles-ci.

Le modèle relationnel, étant basé sur un modèle mathématique, est particulièrement adapté au stockage d'information n-aires, à l'analyse (à travers une modélisation dimensionnelle par exemple) et à la mise à jour des données. Les moteurs de bases de données relationnelles ont été éprouvés avec le temps et proposent aujourd'hui divers mécanismes d'intégrité des données, d'indexation et de tolérance de pannes qui permettent d'améliorer le stockage et l'interrogation des données. Diverses tentatives récentes montrent que la technologie développée autour des modèles relationnels reste très performante, aussi, diverses tentatives de modélisations plus expressives (EAV, dimensionnel, dimensionnel-EAV) ont vu le jour.

		Relationnel	Arbres	Graphes
Stockage d'information	Moteurs	MySQL, Oracle	eXist, XMLDB	Virtuoso, Sesame
	Relations n-aires	+++	++	+
	Robustesse	+++	++	++
	Relations sémantiques	-	-	+++
	Vocabulaires	+	+	++
Recherche d'information	Languages	SQL	XQuery, XPath	SPARQL
	Rapidité	+++	+	+
	Inférence	-	-	+
Echange d'information	Formalisme	CSV	XML	SPARQL
	Fédération	-	+	+++
	Sémantique	-	-	++

FIGURE 2.12 – Tableau récapitulatif des propriétés des structures de stockage d'information.

Il est cependant évident aujourd'hui, que ce modèle est limité pour décrire les relations qui existent entre les objets d'un domaine. Les relations sémantiques sont explicites dans les graphes (autrefois graphes conceptuels) basés sur RDF, ce qui permet la mise en oeuvre de mécanismes d'inférence (SPARQL permet d'interroger par la relation). Nous pensons que les relations entre les objets doivent être explicitement formalisées et stockées avec les données, particulièrement dans le domaine de l'information biomédicale (au regard de la complexité de cette information). Le modèle relationnel, en l'état, ne permet pas de persister cette information, et son langage d'interrogation (SQL) n'offre pas les méthodes explicites d'utilisation de celle-ci lorsque des tentatives ont été faites.

Depuis l'introduction des réseaux sémantiques par Liday dans les années 70. Dans [Roussopoulos 1975], l'idée d'utiliser les réseaux sémantiques pour gérer des données fût émise. Cependant IBM et le modèle relationnel de Codd fût industriellement mieux accepté. Les graphes connaissent cependant un essor important grâce à l'Internet. La configuration en constellation du réseau et les informations véhiculées sur ce réseau doivent pouvoir s'organiser si nous voulons qu'une certaine intégrité soit préservée lorsqu'une information est visible sur celui-ci. Le W3C [14] est aujourd'hui l'organisme international qui produit des standards et des spécifications qui visent à transformer l'Internet comme un espace géant de stockage d'information (LinkedData [15] sur lequel les données seraient porteuses de leur propre sémantique et seraient, dans une certaine mesure, interprétables par l'homme et la machine

14. World Wide Web Consortium : http://www.w3.org
15. http://linkeddata.org/

[Berners-Lee 2001]. Cette accélération de l'adoption des graphes (ou des triplets) comme structure de stockage nous pousse à la question de leur utilisation dans le cadre du stockage d'information biomédicales mais aussi de leur partage. En effet, les capacités de partage de ce type de structures sont d'une part plus étendues (sémantique, fédération, vocabulaires) et d'autre part certainement plus adaptées au stockage et au traitement de l'information biomédicale puisque la relation entre deux concepts est aussi importante que les concepts eux-mêmes dans leur usage.

La lecture du tableau 2.12 montre aussi qu'il y a une relation inverse entre la complexité de la structure de stockage, et sa capacité à être interrogée. Plus le modèle logique sera capable d'exprimer la complexité de l'information stockée, plus il est difficile d'interroger de grands volumes de données avec de bonnes performances. C'est une problématique actuelle forte, et nous proposerons des solutions méthodologiques pour le contourner.

2.3 Modèles de connaissance : Référentiels, Terminologies et Ontologies

Une des spécificités du monde médical est la richesse de la connaissance à manipuler. UMLS [16] contient 1 million de concepts biomédicaux et 5 millions de termes. Il n'est pas rare d'avoir plusieurs termes pour définir le même concept et vice versa. La même maladie peut être désignée par des noms ou des expressions différentes (synonymie), le même terme peut avoir un sens différent suivant le contexte ou le locuteur (polysémie). Cette situation rend la numérisation de l'information médicale difficile, c'est la raison pour laquelle la discipline s'est intéressée aux systèmes de codage et de structuration de l'information issue de l'ingénierie des connaissances. La connaissance médicale évolue constamment. Il est donc logique (afin de garantir une pérennité des systèmes d'informations biomédicaux) de séparer l'information médicale en plusieurs ensembles : les données, les terminologies et la connaissance. Alors que les données sont des ensembles relatifs aux événements réels arrivant au patient (par exemple), les terminologies formaliseront l'information relative aux termes utilisés pour stocker cet événement. Enfin, les ontologies représenteront les connaissances associées au domaine de l'événement. La connaissance peut être stockée grâce à des modèles d'informations de données logiques vus précédemment. Cependant, comme nous l'avons vu, les modèles souvent adaptés pour le stockage de l'information numérique et structurée ne le sont pas forcément pour l'information littérale. De plus,

16. Unified Medical Language System : ressource sémantique visant à stocker toutes les ressources termino-ontologiques biomédicales - http://www.nlm.nih.gov/research/umls/

nous allons le voir, la notion de relation sémantique entre les éléments d'information est, dans le cas de la connaissance, primordial. Nous présenterons dans cette section les différents modèles de représentation proposés par la communauté de l'ingénierie des connaissances. Nous aborderons particulièrement les notions de terminologie et de classifications telles qu'on peut les voir dans les systèmes d'informations actuels. Nous parlerons ensuite d'ontologies d'abord en présentant succinctement ce que sont les ontologies, puis en présentant comment l'information s'organise au sein d'un système d'organisation de la connaissance. Nous remarquerons alors qu'il existe divers niveaux de connaissance et aussi de modélisation de la même réalité qu'il convient, peut être, d'organiser.

2.3.1 Définitions

Puisque nous abordons les notions d'organisation de la connaissance, commençons par définir de manière formelle, avec l'aide du Dictionnaire Robert de la langue française et de wikipedia ce que sont les différentes formes d'organisation connues :

- Un dictionnaire est un "recueil d'unités signifiantes de la langue (mots, termes, éléments...) rangées dans un ordre convenu, qui donne des définitions, des informations sur les signes".
- Une nomenclature est "1- l'ensemble des termes employés dans une science, une technique, un art, méthodiquement classés, 2- l'ensemble des formes (mots, expressions, morphèmes) répertoriées dans un dictionnaire, un lexique et faisant l'objet d'un article distinct".
- Un thésaurus est un "répertoire alphabétique de termes normalisés pour l'analyse de contenu et le classement des documents d'information".
- Une classification est "l'action de distribuer par classes, par catégories le résultat de cette action". En biologie, les organismes sont classés en règnes, embranchements, classes, ordres, familles, tribus, genres, espèces, sous-espèces, variétés, races et formes.
- Une terminologie est "l'ensemble de termes, rigoureusement définis qui sont spécifiques d'une science, d'une technique, d'un domaine particulier de l'activité humaine".
- Une ontologie est "l'ensemble structuré des termes et concepts représentant le sens d'un champ d'informations, que ce soit par les métadonnées d'un espace de noms, ou les éléments d'un domaine de connaissances. L'ontologie constitue en soi un modèle de données représentatif d'un ensemble de concepts dans un domaine, ainsi que des relations entre ces concepts. Elle est employée pour raisonner à propos des objets du domaine concerné".

Ces différentes natures d'organisation sont toujours construites en fonction d'un domaine d'étude, tout comme un modèle d'information. Nous détaillerons dans le chapitre suivant les différentes formes expressives que peuvent prendre les ontologies en fonction du langage utilisé.

2.3.2 Terminologies et Classifications

"La terminologie, considérée comme une science, s'intéresse au recensement des concepts d'un domaine et des termes qui le désignent pour faciliter l'échange de connaissances dans une langue et d'une langue à l'autre." [Baneyx 2007]. Une terminologie implique la normalisation des termes d'un domaine afin de pouvoir les organiser les uns par rapport aux autres [Zweigenbaum 1999]. Le principal intérêt d'une terminologie est de réduire l'ambiguité qui existe entre les termes d'un domaine et de pouvoir donc mieux partager de l'information. [Charlet 2002] définit une classification comme l'action de distribuer par classes et par catégories les concepts d'un domaine. C'est une répartition systématique par classes ou catégories ayant des caractères communs dans un contexte précis [Bourigault 2004]. La structure et la profondeur de la classification dépendent de l'objectif du concepteur.

La classification internationale des maladies (CIM) qui vise à classer les maladies en fonction de leur *nature* ou de *l'endroit anatomique* est un exemple de classification. La liste suivante représente la table analytique de la CIM.

```
1 Classification des maladies et traumatismes
1.1 A00-B99 - Certaines maladies infectieuses ou parasitaires
1.2 C00-D48 - Tumeurs
1.3 D50-D89 - Maladies du sang et des organes hématopoïétiques et
certains troubles du système immunitaire
1.4 E00-E90 - Maladies endocriniennes, nutritionnelles et métaboliques
1.5 F00-F99 - Troubles mentaux et du comportement
1.6 G00-G99 - Maladies du système nerveux
1.7 H00-H59 - Maladies de l'œil et ses annexes
1.8 H60-H95 - Maladies de l'oreille et de l'apophyse mastoïde
1.9 I00-I99 - Maladies de l'appareil circulatoire
1.10 J00-J99 - Maladies de l'appareil respiratoire
1.11 K00-K93 - Maladies de l'appareil digestif
1.12 L00-L99 - Maladies de la peau et du tissu sous-cutané
1.13 M00-M99 - Maladies du système ostéo-articulaire, des muscles et du
tissu conjonctif
```

1.14 N00-N99 - Maladies du système génito-urinaire
1.15 O00-O99 - Grossesse, naissance et la période puerpérale
1.16 P00-P96 - Certains états qui trouvent leur origine dans la période périnatale
1.17 Q00-Q99 - Malformations congénitales, déformations et anomalies chromosomiques
1.18 R00-R99 - Symptômes, signes et observations cliniques ou de laboratoire anormales, non classées ailleurs
  1.18.1 (R00-R09) Symptômes et signes impliquant le système circulatoire et respiratoire
  1.18.2 (R10-R19) Symptômes et signes impliquant le système digestif et l'abdomen
  1.18.3 (R20-R23) Symptômes et signes impliquant la peau et le tissu sous-cutané
  1.18.4 (R25-R29) Symptômes et signes impliquant le système nerveux central et l'appareil locomoteur
  1.18.5 (R30-R39) Symptômes et signes impliquant le système urinaire
  1.18.6 (R40-R46) Symptômes et signes impliquant la cognition, la perception, l'humeur et le comportement
  1.18.7 (R47-R49) Symptômes et signes impliquant la parole et la voix
  1.18.8 (R50-R69) Symptômes généraux et signes
  1.18.9 (R70-R79) Valeurs sanguines anormales, sans diagnostic
  1.18.10 (R80-R82) Valeurs anormales dans l'urine, sans diagnostic
  1.18.11 (R83-R89) Valeurs anormales dans l'examen d'autres liquide corporel, substance ou tissus, sans diagnostic
  1.18.12 (R90-R94) Valeurs anormales dans l'imagerie médicale ou une épreuve fonctionnelle, sans diagnostic
  1.18.13 (R95-R99) Causes de mortalité liées à une maladie et inconnus
1.19 S00-T98 - Traumatisme, intoxication et certaines autres conséquences de causes externes
1.20 U00-U99 - Codes pour certains cas spéciaux
1.21 V01-Y98 - Causes externes de morbidité et de mortalité
1.22 Z00-Z99 - Facteurs influençant l'état de santé et le contact avec les services de santé

Les relations entre les différents concepts d'une classification peuvent être de type spécification-généralisation (de type "est-un" ou "is-a", par exemple : "escherichia coli est une bactérie") ou de type partition ("fait-partie-de" ou "part-of", comme

dans : "le coeur fait partie de l'appareil circulatoire"). Les hiérarchies mono-axiales ont une seule racine et doivent couvrir l'ensemble du domaine d'étude sans doublons, ce qui est très compliqué à mettre en oeuvre, voire impossible dans beaucoup de classifications de la médecine. C'est pourquoi des classifications multi-axiales ont été proposées rapidement, comme le MeSH [17] ou l'ATC. Il est à noter que les terminologies visent tout d'abord à organiser des listes de termes. Elles peuvent mettre en oeuvre, en plus de l'organisation hiérarchique proposée, des définitions. Cependant, la relation entre deux termes est généralement d'une forme simple ("is-a"). SKOS [18] propose un jeu de relations plus étendues comme les BTNT [19] afin de pallier certaines problématiques de modélisation. L'utilisation d'une terminologie apportera à un système d'information un outil de contrôle de vocabulaire de son domaine, tout en lui permettant d'utiliser les relations entre les termes dans le cadre de recherche d'information par exemple. Cette limitation de la relation entre les concepts permet de garantir la décidabilité et la performance du système d'information. L'ontologie permet, elle, l'utilisation de relations plus riches et d'organiser la connaissance avec plus d'expressivité.

2.3.3 Ontologies

Le terme d'ontologie dans le domaine des sciences de l'information a été repris et diffusé notamment par [Gruber 1993] qui définira une ontologie comme "une spécification partagée d'une conceptualisation". Ce terme apparaît en philosophie au XVIIe siècle après avoir été introduit par Aristote comme une partie de la philosophie qui a pour objet "l'être en tant qu'être", autrement dit, la science de l'être. L'ingénierie des connaissances, a depuis les années 90 contribué à l'émancipation du terme et plusieurs définitions seront proposées. Les ontologies partagent des propriétés avec les taxonomies, les classifications ou bien les thesaurus, cependant, elles ne se limitent pas simplement à la définition de classes et aux relations de subsomption. Pour spécifier une conceptualisation, il sera nécessaire de définir les axiomes qui pourront contraindre l'interprétation des termes définis. Charlet propose [Charlet 2002] une définition affinée de ce qu'est une ontologie : "Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts - e.g. entités, attributs, processus -, leurs définitions et leur interrelations. On appelle cela une conceptualisation." [...] "Une ontologie peut prendre différentes formes mais elle incluera nécessairement un vo-

17. Medical Subject Headings
18. Simple Knowledge Organisation System
19. Narrower than - Broader than

cabulaire de termes et une spécification de leur signification." [...] "Une ontologie est une spécification rendant *partiellement* compte d'une conceptualisation". Cette dernière définition précise les précédentes et introduit un terme important, à notre sens : "partiellement". En effet, ce terme met en avant la notion d'incomplétude d'une ontologie, quelle qu'elle soit. Particulièrement dans le domaine biomédical, il est essentiel de préciser qu'une ontologie ne peut (ni ne doit) décrire toute la médecine de manière consensuelle et partagée. La complexité d'une telle ontologie (même si l'on trouvait le consensus nécessaire entre les experts de ce domaine scientifique) la rendrait inutilisable par l'homme et par la machine à cause de la combinatoire et de l'indécidabilité que celle-ci engendrerait.

2.3.3.1 Anatomie d'une ontologie

Les concepts

Les objets matériels (par exemple : un comprimé de médicament) ou immatériels (une quantité) du monde réel que nous voulons décrire dans une ontologie sont modélisés sous forme de concepts (parfois nommés classes)[Uschold 1995]. Un concept est divisible en 3 notions de natures différentes qui puisent leur définition dans l'approche sémiotique qui s'inspire du triangle aristotélicien : le symbole, le signifié et le signifiant.

Aristote défini la théorie de la signification et des symboles. Il dit : "Il n'est pas possible d'apporter dans la discussion les choses elles-mêmes, mais, au lieu des choses, nous devons nous servir de leurs noms comme de symboles" (Réfutations Sophistiques, 1, 165a).

L'approche sémiotique, qui défini la science de l'étude des signes, propose de nous intéresser à la manière dont nous représentons, de manière formelle ou conceptuelle, les concepts (ou signes) qui nous entourent. On distinguera deux grands courants dans l'approche sémiotique. Tout d'abord, le courant linguistique où Ferdinand de Saussure définira un concept comme ayant un signifiant et un signifié [de Saussure 1916]. C'est à dire que chaque concept du monde pourra être défini par un terme (signifiant) et par son sens (signifié). Ensuite celui de Peirce, un logicien, qui propose un modèle de représentation différent de celui de de Saussure. Il sera à l'origine du triangle sémiotique de Peirce où il décrit un concept suivant 3 axes : le "representamen", l'"interpretant" et l'"object". Ces deux courants sont aujourd'hui réunis sous le concept de la sémiotique, le langage étant selon Claude Lévi-Strauss (Lévi-Strauss, 1972), le système sémiotique par excellence. La figure 2.13 représente le triangle aristotélicien et le triangle sémiotique.

Les relations

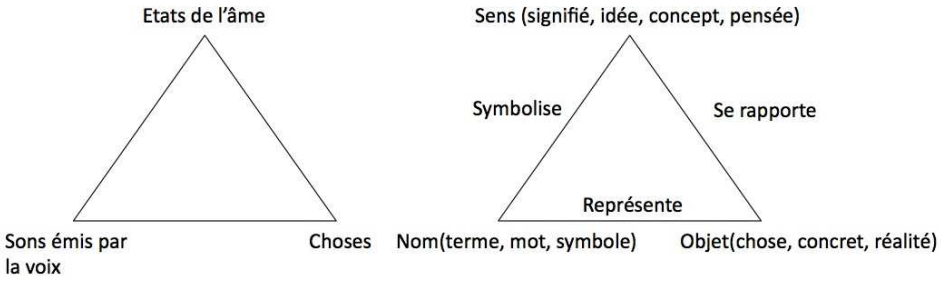


FIGURE 2.13 – Le triangle aristotélicien (gauche) et le triangle sémiotique (droite)

Les relations permettent de lier un ou plusieurs concepts dans une ontologie. Les concepts simples (primitifs) peuvent être reliés entre eux pour définir un concept défini. Les concepts peuvent être structurés en hiérarchie implicite ("is-a") dans le cas d'une ontologie suivant la logique de description [20] qui permettra la mise en oeuvre de certains raisonneurs sur l'ontologie, ou ils peuvent être structurés avec des relations plus ouvertes et non limitées par une logique particulière entre les concepts, les individus et les relations, ce qui est le cas d'OWL Full [21]. La relation de subsomtion ("is-a") est définie de manière implicite. Par exemple un concept père C_1 subsume un concept fils C_2 si toute propriété sémantique de C_1 est propriété sémantique de C_2 et si C_2 est plus spécifique que C_1.
La figure 2.14 décrit les relations de subsomption simples pour décrire l'homme et la femme dans un contexte de classification du vivant. Un être vivant définit tout ce qui vit, il représente un concept non primitif, tout autant que le concept d'humain, la relation qui lie les deux est tout à fait correcte. Un humain sera toujours un être vivant. Un femme est toujours un être vivant, tout autant qu'un concept de genre féminin. De même, un animal peut être mâle ou femelle (ou les deux, ce qui n'est pas représenté ici). Un homme est généralement toujours du genre masculin (ou mâle). Nous remarquons ici que le sens de la relation est important. En effet, si tous les hommes sont des mâles, tous les mâles ne sont pas nécessairement des hommes. On définira la relation de subsomption de plusieurs manières :

20. Les logiques de description aussi appelées logiques descriptives (LDs) sont une famille de langages de représentation de connaissances qui peuvent être utilisés pour représenter la connaissance terminologique d'un domaine d'application d'une manière formelle et structurée. Le nom de logique de description se rapporte, d'une part à la description de concepts utilisée pour décrire un domaine et d'autre part à la sémantique basée sur la logique qui peut être donnée par une transcription en logique des prédicats du premier ordre. La logique de description a été développée comme une extension des frames et des réseaux sémantiques, qui ne possédaient pas de sémantique formelle basée sur la logique. *Wikipédia*

21. http://www.w3.org/TR/owl-ref/

- Définition intentionnelle : un concept C_1 subsume un concept C_2 si tout individu décrit par C_2 l'est aussi par C_1, c'est à dire si l'ensemble des propriétés d'un individu dont la description est définie par C_2 contient l'ensemble des propriétés spécifiées par C_1.
- Définition extentionnelle : un concept C_1 subsume un concept C_2 si l'ensemble des individus dénotés par C_1 contient l'ensemble des individus dénotés par C_2. Par exemple, le concept "pathologie" subsume le concept "pneumonie".
- Définition logique : un concept C_1 subsume un concept C_2, si être un individu décrit par C_1 implique être un individu décrit par C_2.

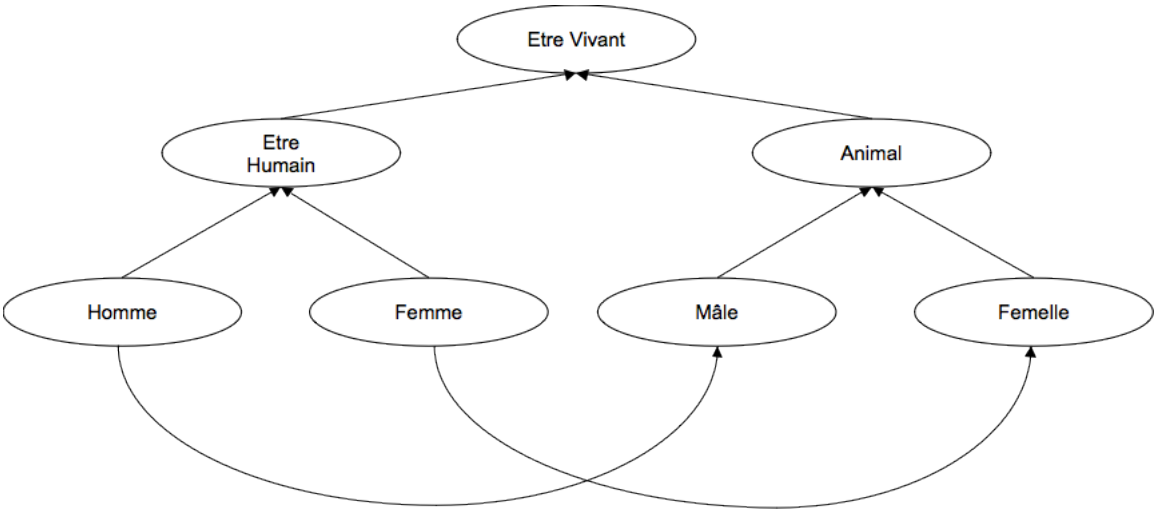


FIGURE 2.14 – Exemple de relation de subsomption

Les relations entre les concepts définissent aussi le type de raisonneurs utilisables sur l'ontologie. Le web sémantique définit ce type d'ontologie grâce au formalisme OWL-DL [Horrocks 2003]. La logique de description divise la connaissance en 2 parties : les informations terminologiques (les concepts et leur relation) et les informations sur les individus (les relations entre les individus). Cette division implique qu'un individu ne peut être un concept et vice versa. Si ce cas devait se présenter, alors une telle ontologie serait formalisée en OWL-Full. Les contraintes d'OWL-DL ont été introduites afin de suivre les travaux effectués en logique de description par les raisonneurs développés alors. Notons que certains raisonneurs sont capables d'inférer sur des ontologies en OWL-Full [De Roo 2002] grâce à l'utilisation de détecteurs de boucles. Quand à OWL-Lite, il sera surtout utilisé dans le cadre de systèmes de classification simples.

Les instances

Les instances d'un concept représentent généralement les individus du concept. Par exemple, l'instance d'un concept de Bactérie peut être Escherichia Coli. Ce n'est cependant pas toujours le cas. En effet, dans le cas d'une ontologie de maladies infectieuses, Escherichia coli sera un concept qui aura des relations avec d'autres concepts tels que par exemple, le commensalisme ou l'infection. De même dans notre exemple 2.14, Pierre est un homme, il peut donc être soit une instance de Homme, ou bien un concept ayant comme terme Pierre et qui représente une personne physique réelle.

2.3.3.2 Natures d'ontologies

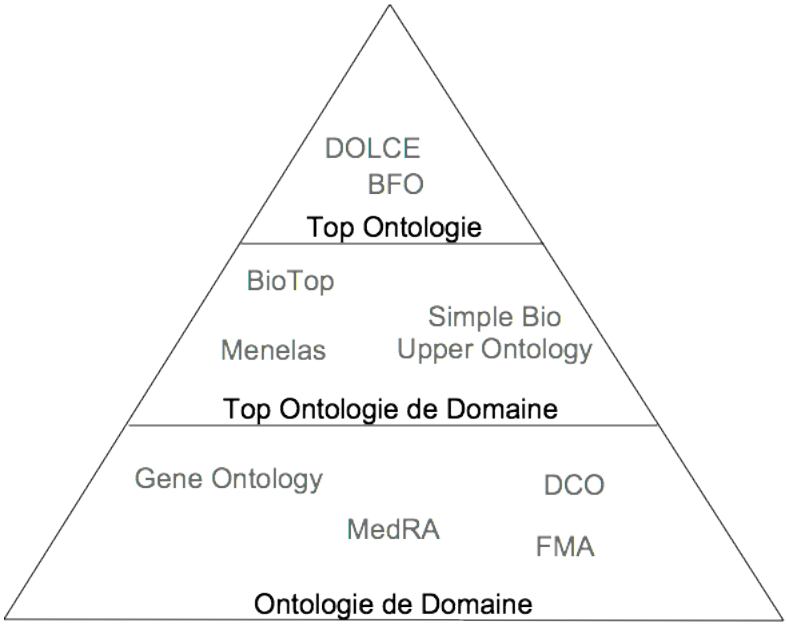


FIGURE 2.15 – Pyramide des niveaux ontologiques. Adaptés depuis Alan Rector.

La littérature définit divers types d'ontologies [Gomez-Perez 2004]. En fonction des objets qu'elles modélisent, de leur degré de granularité et de leurs topologies, les ontologies peuvent être de nature différentes. Les travaux de [Mizoguchi 1995], [Van Heijst 1997] et de [Guarino 1998] vont nous permettre de définir différents types d'ontologie dans cette section. Par exemple, ils proposent une classification en fonction des objets modélisés dans les ontologies en rapport à leur objectif, à savoir :

– Ontologies de domaine [Mizoguchi 1995, Van Heijst 1997]. Ce type d'ontologie permet de spécifier les connaissances associées à un domaine particulier. Elles modélisent et rendent compte du vocabulaire d'un domaine au travers de concepts et de relations modélisées suivant les activités, les théories et les

principes de ce domaine. L'ontologie des maladies infectieuses, des urgences ou bien de la médecine générale sont des exemples d'ontologie de domaine. Les ontologies de ce type sont généralement structurées grâce à des ontologies de haut niveau.

– Ontologies de domaine de haut niveau [Schulz 2006]. Les ontologies de haut niveau permettent de structurer les concepts et relations d'un ou de plusieurs domaines. BioTop, par exemple, est une ontologie de domaine de haut niveau qui vise à aider à l'organisation des concepts du domaine de la biologie et de la santé.
– Ontologies de haut niveau [Ikeda 1997, Guarino 1997]. Les ontologies de haut niveau (par exemple DOLCE [22] ou BFO [23]) sont des ontologies plus philosophiques qui visent à définir les éléments structurant de toute ontologie de domaine.

La figure 2.16 présente un exemple de modélisation entre l'ontologie de domaine GO [24], l'ontologie de domaine de haut niveau BioTop et l'ontologie de haut niveau. Nous remarquons ici les différents éléments constituants de ces 3 types d'ontologies.

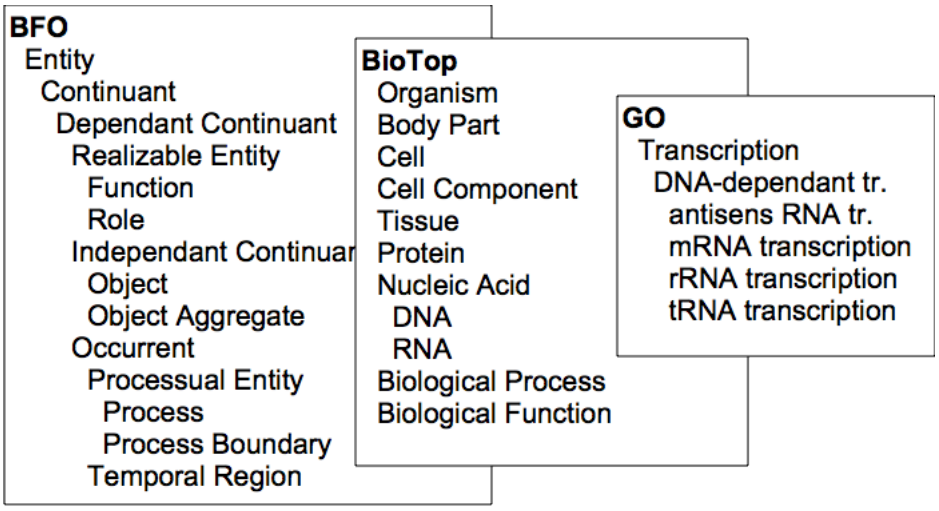


FIGURE 2.16 – Exemple d'ontologie de haut niveau, d'ontologie de domaine de haut niveau et d'ontologie de domaine. Adapté depuis Stefan Schulz.

L'organisation des systèmes de représentation de la connaissance est nécessaire dans un domaine aussi vaste que celui de l'information biomédicale. Nous verrons

22. http://www.loa.istc.cnr.it/DOLCE.html
23. http://www.ifomis.org/bfo
24. Gene Ontology

plus loin comment nous utilisons cette organisation dans le cadre du domaine d'application de notre travail.

2.4 Modèles de qualité

A ce jour, la mise en œuvre d'architectures d'informations pour l'analyse de données est une bonne opportunité pour les institutions de santé d'améliorer la qualité de leurs données, et de replacer leurs données dans leur contexte sémantique [Wisniewski 2003].

Parmi les modèles de traitement de l'information, nous distinguons les modèles de qualité de l'information comme une source de connaissances visant à caractériser des données dans un but de partage de l'information (tout comme nous l'avons vu avec les modèles d'informations ou de connaissances). Nous abordons ici rapidement les méthodes et les outils proposés dans diverses communautés pour aider à mesurer la notion de qualité de l'information. Nous verrons dans le chapitre 6 comment nous proposerons de mettre en oeuvre partie de ces propositions dans une méthode globale de mesure de la qualité de l'information dans un cadre d'interopérabilité.

2.4.1 La qualité de l'information

Nous présentons dans cette section tout d'abord l'état de l'art des mesures proposées dans la littérature en général pour évaluer la qualité de l'information. Nous présentons ensuite les méthodes principales utilisées dans la mise en oeuvre de système d'amélioration de la qualité. Nous aborderons ensuite des exemples de projet qualité et nous nous intéresserons plus particulièrement à ce qui a été proposé dans le domaine de l'information biomédicale.

2.4.1.1 Les mesures de qualité

Devant l'accroissement des quantités de données numérisantes, la notion de qualité des données fût étudiée par les statisticiens vers la fin des années 60 [Fellegi 1969]. Au début des années 90, les sciences de l'information ont commencé à formaliser la problématique de la mesure et de l'amélioration de la qualité des données. L'ISO définit la qualité comme « la totalité des spécificités et des caractéristiques d'une entité qui permettent de satisfaire aux usages implicites et explicites définis » (ISO 8402-1986 Vocabulaire Qualité). Dans cette lignée, [Wang 1998] propose de définir la qualité d'une donnée en fonction de l'usage attendu que l'on en a. Bien que ce concept d'utilité attendue soit assez générique pour définir un principe, il faudra attendre [Redman 1996] pour une caractérisation du concept de qualité des données

suivant 4 dimensions : exactitude, perfection, fraîcheur et uniformité. D'autres facteurs de qualité ont été proposés afin de mesurer la qualité des données en fonction de processus [Naumann 2000] et de leur but [Peralta 2008].

La communauté des entrepôts de données a aussi proposé des approches afin de mesurer, d'améliorer et de surveiller la qualité des données [Weikum 1999] [Berti-Equille 2005]. [Wand 1996] définit les dimensions de qualité de l'information selon les fondements de la sémiotique dans une approche que nous qualifierons de descendante.

En marge de ces travaux, l'ingénierie des modèles apporte des méthodes afin de mesurer la qualité des modèles d'information [Krogstie 2010] [Moody 2003b]. Il est cependant difficile de mesurer la qualité d'expression d'un modèle d'information avec seulement des métriques [Moody 2003a]. Moody définit alors des facteurs de qualité qui servent de support à leur méthode d'évaluation subjective (ou empirique) d'un modèle d'information où un expert note de 0 à 5 chacun des 7 critères suivants : l'exactitude, l'applicabilité, la complétude, l'intelligibilité, l'intégration, la flexibilité et la simplicité.

2.4.1.2 Les processus d'amélioration de la qualité

Afin d'aborder la notion d'amélioration de la qualité de l'information, [Wang 1998] propose une méthodologie basée sur la roue de Deming [Deming 2000] (définir, réaliser, contrôler, agir) nommée TDQM [25] qui définit un processus itératif d'amélioration de la qualité des données.

La roue de Deming utilise la méthode PDCA (Plan-Do-Check-Act) définie de la manière suivante :

- Plan : Planification des objectifs, de ce qu'on va réaliser
- Do : Développer, réaliser, mettre en oeuvre les actions correctives
- Check : Contrôler, vérifier l'atteinte des objectifs
- Act : Agir, réagir, prendre des mesures préventives selon les résultats précédents

Ce cycle est itératif et chaque itération est censée apporter une amélioration de la qualité selon les critères définis.

L'ingénieur Mikel Harry défini les bases de la méthodologie "six sigma" en se basant sur la philosophie de celle de Deming. Cette méthode a d'abord été appliquée pour la maîtrise statistique des procédés industriels avant d'être élargie à tous types de processus. La lettre Sigma désigne l'écart-type et donc "Six Sigma" consiste à faire en sorte que tous les éléments issus du processus étudié soient compris dans un

25. Total Data Quality Management

intervalle s'éloignant au maximum de 6 sigma par rapport à la moyenne générale des éléments issus de ce processus. La méthode se base sur 5 étapes DMAAC (Définir, Mesurer, Analyser, Améliorer et Contrôler).

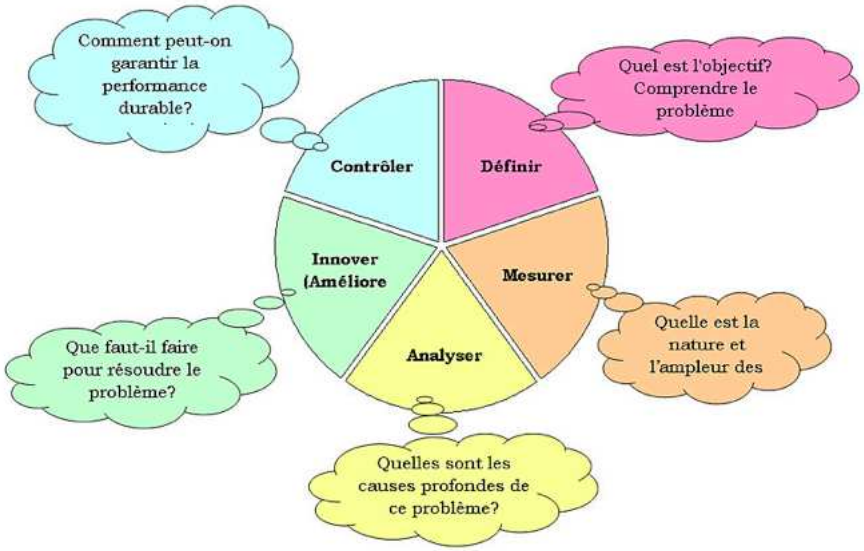


FIGURE 2.17 – Le processus d'amélioration de la qualité Six Sigma.

2.4.1.3 Exemples de projets qualité

Le projet QUADRIS s'inspire du méta-modèle développé dans le projet DWQ (DataWarehouse Quality) [Vassiliadis 1999] et de l'approche Goal-Question-Metric [26] [Basili 2002] et s'appuie sur l'utilisation de patrons. Ce projet a été conçu dans le but d'offrir un cadre d'évaluation de la qualité dans les systèmes d'informations multisources. Dans ce projet, chaque dimension peut être déclinée en plusieurs facteurs (par exemple de précision syntaxique d'une liste de termes).

Dans le domaine de la santé, des travaux spécifiques ont aussi été proposés du fait du besoin croissant d'utiliser les données du dossier patient à des fins d'analyses épidémiologiques ou de santé publique. Néanmoins, l'utilisation de ces données est souvent freinée, là aussi, par la mauvaise qualité des données [Goldberg 2008]. Les causes de défaillance de qualité dans les données de santé ont été classifiées en deux types d'erreurs : systématiques et hasardeuses [Arts 2002] aux différents niveaux du processus de saisie. [Kerr 2007] présente un cadre de mesure de la qualité des données basées sur les recommandations du CIHI [27]. Cette étude débouche sur la classifica-

26. GQM est une approche permettant de spécifier des buts de qualité qui sont déclinés en un ensemble de questions auxquelles on associe une mesure de qualité

27. The Canadian Institute for Health Information

tion de 69 critères de qualité regroupés dans 24 caractéristiques qui se subdivisent en 6 dimensions : précision, ponctualité, comparabilité, utilisabilité, pertinence et sécurité.

2.4.2 Synthèse

Nous n'abordons pas toute la littérature sur la qualité de l'information, car elle est très large. Nous nous sommes concentrés sur ce qui était proposé et qui était, à priori, utilisable et adapté à notre problématique de qualification de l'information échangée. La plupart des mesures et méthodes proposées tentent d'être généralisables. Aussi, il a été difficile d'estimer quelles mesures étaient utiles dans notre cadre. De même, chaque mesure n'est pas spécifiquement utile pour tous les types d'information. Le modèle d'information par exemple, doit, pour être interopérable, être facilement compréhensible par l'homme, et la machine. Les terminologies ne se mesureront pas suivant les mêmes axes, etc. Nous proposerons dans le chapitre 6 un cadre de classification de ces mesures (nombreuses) en fonction de la nature de l'information mesurée.

2.5 Synthèse

Les modèles d'information sont adaptés à la gestion de données dans les bases de données, mais ils souffrent cependant d'une lacune : l'expression de la relation, ce qui rend implicite une partie de l'information nécessaire à l'utilisation de ces bases de données. Les modèles en graphe ou en réseau permettent d'exprimer la sémantique de la relation et de l'interroger. La capacité de raisonnement qu'offrent d'ailleurs ces modèles (RDF, Associatif) est rendue possible grâce à la relation elle-même. La relation apporte un contexte sémantique. Les modèles d'information actuels (relationnels le plus souvent) doivent être capables d'évoluer vers des représentations plus expressives. Mais nous l'avons vu, le gain d'expressivité peut être inversement corrélé à la facilité et à la performance d'interrogation du modèle, mais pas toujours. En effet, il nous semble que le couplage de la connaissance et des données dans le même modèle de représentation (RDF) peut permette de faciliter l'interrogation des données en gardant une expressivité riche.

Nous verrons dans la suite de cette thèse une proposition de méthode d'intégration des données et de la connaissance tout en gardant les avantages de chaque type de modélisation (relationnelle et graphe). Il est vrai que l'internet et les projets tels que le Linked Open Data nous poussent à revoir les paradigmes de l'intégration de données à grande échelle pour l'analyse de données en particulier. La modélisation

RDF en triplets permet, avec un jeu de relations contrôlé, d'exprimer la plupart des modèles actuels, et bien plus (RDF-S). Nous pensons que nous arrivons, historiquement, à un point d'inflexion où modèles de données et modèles de connaissances tentent de se rejoindre.

Dans le chapitre suivant nous aborderons les problématiques de partage d'information. Forts de la connaissance que nous avons sur les modèles d'information, les modèles de connaissance et les modèles de qualité de l'information, nous aborderons la problématique de partage de l'information au niveaux technique, syntaxique et sémantique. Nous pensons que le partage d'information doit se faire avec le moins d'ambiguïté possible afin que les deux locuteurs (dans ce cas des systèmes informatiques) soient capable d'interpréter le sens de celle-ci. Les modèles présentés dans ce chapitre nous aideront à apporter du sens à cet échange. La réflexion autour de l'interopérabilité sémantique nous mènera naturellement à la vision du web sémantique, puis, à l'étude de la mise en oeuvre d'une plateforme d'interopérabilité sémantique pour des données biomédicales dans le cadre du projet européen DebugIT.

2.6 Discussion

La modélisation des connaissances que nous avons d'un domaine peut aider d'une part à mieux comprendre le domaine, et peut permettre d'autre part, grâce à des outils mathématiques, statistiques ou algorithmiques, de découvrir de nouvelles connaissances dans le domaine. L'outil informatique encourage les concepteurs de systèmes d'information à structurer et à formaliser l'information médicale dans des systèmes d'information générant toujours plus de données. Cette organisation de l'information conduit forcément à un choix : plus on est expressif, plus les performances de gestion se dégradent, et inversement. Or la gestion de l'information médicale demande de l'expressivité. En effet, la complexité et le volume du vocabulaire médical (plus d'un million de termes dans UMLS) demandent la mise en oeuvre de terminologies et d'ontologies. Mais peut-on tout modéliser ?
D'après Heidegger,"La science ne pense pas.". Cet axiome exprime, entre autres, le fait que la science ne peut se penser elle-même. Or une des propriétés de la pensée est de pouvoir s'observer, se voir en train de penser. La mathématique ne peut prouver la mathématique. Gödel (1931) nous dit : "Aucun langage rationnel ne peut rendre compte totalement du monde". Pour lui, il y a une certaine limite inhérente à tout système de terminologie (la langue étant le système terminologique par excellence), c'est sa capacité à représenter la totalité du monde, mais par extension nous dirons d'un domaine particulier. Nous avons observé cette problématique dans le cadre de

cette thèse et nous y reviendrons. Cependant, il est important de garder à l'esprit que même le langage ne peut représenter la totalité du monde, le langage étant le système le plus expressif que nous connaissons. Par extension, il est difficile de demander à un modèle d'information, à une ontologie, d'atteindre une expressivité parfaite. C'est pourquoi une question se pose aujourd'hui autour des formalismes et de la logique de certaines modélisations. En effet, certains groupes de pensées dans la communauté de l'informatique médicale tendent à penser que la modélisation en hypothèse de monde fermé [28] est préférable afin d'assurer la cohérence des décisions médicales prises lors de l'utilisation du système d'aide à la décision par exemple. Il est cependant intéressant d'observer que, puisque toute la médecine n'est modélisable dans aucun système où toutes les hypothèses seraient connues, ne serait-il pas alors intéressant pour la médecine de mieux appréhender ces mondes ouverts ?

28. Un monde fermé, en logique, est un monde où tout ce qui n'est pas connu est faux, alors que le monde ouvert reconnaît l'existence de l'inconnu. Par exemple : sachant que "François est un citoyen de France". La question "Est ce que Margot est une citoyenne de France ?" aurait comme réponse "Non" dans l'hypothèse d'un monde fermé, et la réponse "Ne sait pas" dans l'hypothèse d'un monde ouvert.

Chapitre 3

Partage d'information

"Face à la croissance explosive des techniques de communication de l'information, les capacités de notre cerveau d'acquérir, de stocker, d'assimiler et d'émettre de l'information sont restées inchangées." - Pierre Joliot

Sommaire

3.1 Introduction 50
3.2 Interopérabilité 51
3.3 Modèles d'intégration de données 53
3.3.1 Intégration centralisée et persistante 55
3.3.2 Intégration centralisée ou décentralisée non persistante 57
3.3.3 Intégration à la volée ou "mashup" 58
3.4 Le Web Sémantique 59
3.4.1 Langages de représentation 61
3.4.2 Langages de règles 63
3.4.2.1 SWRL / Semantic Web Rule Language 64
3.4.2.2 Turtle N3 64
3.4.3 SPARQL - Standard Protocol and RDF Query Language . . 65
3.4.4 Logiques et Raisonneurs 66
3.5 Conclusion 68

The Web was designed as an information space, with the goal that it should be useful not only for human-human communication, but also that machines would be able to participate and help users communicate with each other. A major obstacle to this goal is the fact that most information on the Web is designed solely for human consumption. Computers are better at handling carefully structured and well-designed data, yet even where information is derived from a database with well-defined meanings, the implications of those data are not evident to a robot browsing the web. More information on the web needs to be in a form that machines can

'understand' rather than simply display.- Tim Berners-Lee, in Nature, 2001. [1].

Les modèles structurant l'information et les modèles structurant la connaissance offrent une formalisation nécessaire au partage de l'information. Nous avons cependant remarqué que le partage d'information est dépendant d'une part de la modélisation de l'information et de la formalisation du modèle d'information, mais aussi du contenu de l'information partagée et véhiculée dans ces modèles d'information. La qualité de l'information partagée est aussi un facteur de facilité de partage et nous proposerons dans ce cadre une méthodologie d'évaluation dans le chapitre 6. Forts de la connaissance associée aux modèles d'information et aux modèles de représentation de la connaissance d'un domaine, nous allons dans ce chapitre faire l'inventaire des modèles de partage d'information. Nous remarquerons que ces modèles ont évolué et connaissent aujourd'hui un nouvel essor avec l'avénement du web de données qui définit de nouveaux standards et de nouveaux paradigmes pour le partage d'information.

3.1 Introduction

Pour partager de l'information, il faut connaître les propriétés de celle-ci. Elle se décline généralement suivant 3 axes, la localisation, la structure et le sens. En sciences de l'information, nous parlons d'interopérabilité. On distingue généralement la problématique d'intéropérabilité de celle de l'intégration de données. Intégrer des données, c'est être capable de relier des données provenant de diverses sources d'information dans un but précis, alors que l'interopérabilité est plus relative à la capacité qu'ont deux systèmes communicants (homme ou machine) à se comprendre. Pour bien comprendre cette problématique de la communication, revenons à une conception humaine de celle-ci. La communication d'information entre êtres humains est possible pour plusieurs raisons :

- le son émis par la voix est entendu par les deux acteurs d'une conversation (où par texte interposé) (nous parlons ici de données en science de l'information),
- le sens des sons (ou des mots) est compréhensible par les deux intervenants et leur rappelle une référence qu'ils ont mémorisée (nous parlons ici en partie, de syntaxe),

1. http://www.nature.com/nature/debates/e-access/Articles/bernerslee.htm

– l'addition de ces sons (ou des mots en phrase) leur permet de comprendre le contexte d'utilisation de ces sons (ou mots) et donc d'en déduire le véritable sens de la communication (nous parlons ici de sémantique).

Evidemment, chaque personne ayant son référentiel d'interprétation, il est difficile d'être absolument certain que le sens soit réellement partagé. En sciences sociales, il est montré que c'est lors de l'échange du message entre les deux parties, que l'alignement sémantique entre les deux référentiels se fait. Le va et vient de messages permet, de manière implicite, d'ajuster les discours et les référentiels. Eventuellement, chaque partie mettra à jour son référentiel d'interprétation en fonction de la connaissance nouvellement acquise, mais pas toujours. Parfois il faudra juste mettre en oeuvre un lien de synonymie entre 2 concepts de ces référentiels, car ils veulent dire la même chose dans un contexte précis. Mais parfois on ne trouvera pas de terrain d'entente, mais le message sera passé. Afin de pallier cette problématique en informatique, il est souvent proposé de standardiser la communication. Par exemple, standardiser les méthodes de transport et d'accès aux données (XML, HTTP, etc.), ou bien standardiser les méthodes de structuration des données (modèles d'information), voir standardiser les données elles-mêmes (vocabulaires). Récemment, le sens même des données a été standardisé (ontologies standard), autant qu'il est possible de le faire. Nous aborderons dans ce chapitre diverses méthodes de partage d'information actuellement présentées dans diverses communautés (bases de données, ingénierie des modèles, web sémantique) afin de mieux cerner le domaine et ses limites. Nous aborderons d'abord la problématique de l'intéropérabilité, puis nous nous intéresserons à la problématique plus précise de l'intégration de données.

3.2 Interopérabilité

Frydman disait : "Information Silos are Everywhere. But so is the Internet" [2]. Le problème de la communication entre systèmes d'information n'est pas nouveau. Beaucoup de recherches sont en cours à ce sujet et continueront de l'être. Des nouvelles technologies émergent continuellement, des nouveaux langages de programmation, des nouveaux moteurs de bases de données, des nouveaux schémas logiques et conceptuels, etc. L'Internet joue actuellement un rôle fédérateur afin de pousser les communautés à standardiser l'information et les méthodes d'accès à celle-ci. La dernière grande évolution de l'intéropérabilité en date est un sous-projet du web sémantique : le linked data [Berners-Lee 2009]. Ce projet initié par Tim Berners-Lee vise à rendre interopérable sur la toile non pas des documents qu'on relie grâce à des

2. dans e-patient.net en Novembre 2008

liens hypertextes, mais des données qu'on reliera grâce à leur sens (autrement dit, leur sémantique). Ce projet est aussi pharaonique que l'était le world wide web à ses débuts il y a 20 ans. Cet initiative met en avant des problèmes d'interopérabilité qu'on ne peut résoudre simplement avec des modèles d'information ou des vocabulaires standards. La sémantique est une nouvelle piste sérieuse dans ce domaine.

La problématique de l'intéropérabilité de données sémantiques se pose dans un contexte général d'intéropérabilité qui est divisé en cinq couches dans [Gorman 2006].

- Technique : Concerne l'accès technique aux différents types de stockage de données. Les données opérationnelles peuvent être en texte libre, en format XML, dans des bases de données. Les modèles de données varient d'une source à l'autre, depuis du texte libre jusqu'à des systèmes normalisés grâce à HL7 ou OpenEHR (voir chapitre 4).
- Syntaxique : Les données sources sont contraintes par un vocabulaire standard. Les termes sont cependant souvent codés en texte libre, de manière abrégée ou avec des erreurs.
- Sémantique : Même si les données sont accessibles et partagent une partie de leur vocabulaire, seule une représentation formelle de chaque source de données et l'annotation de ces sources à une représentation formelle du domaine permet une interopérabilité sémantique des données.
- Pragmatique / Dynamique : L'usage de l'information et son échange. La connaissance peut être échangée.
- Conceptuelle : Une vue partagée du monde est établie à un niveau épistémologique.

Il est intéressant de noter les similitudes de cette approche avec la pile du web sémantique. La première version (3.1) de celle-ci fût introduite dans (Tim Berners-Lee, 2003). En effet, cette pile peut se découper suivant les mêmes niveaux conceptuels d'interopérabilité. Les couches techniques et syntaxiques sont implémentées grâce aux URI, XML et RDF ; la couche sémantique grâce à OWL et RDF Schema ; la couche pragmatique grâce aux règles et la couche conceptuelle grâce aux preuves et à la vérité.

Dans le cadre de notre problématique d'échange et d'intégration de données biomédicales nous nous limiterons aux trois premières couches d'interopérabilité : technique (réseau, couche d'accès logique aux données, APIs), syntaxique (type de données, terminologie) et sémantique (sens) pour la mise en oeuvre d"outils d'intégration de données dans le cadre du projet DebugIT.

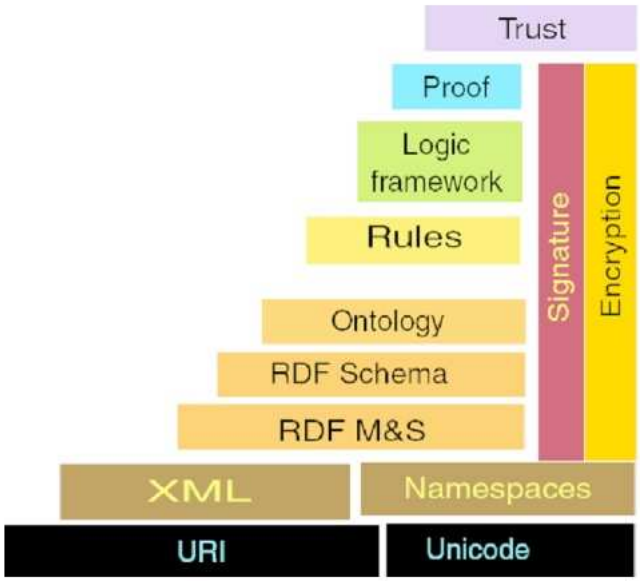


FIGURE 3.1 – La pile du web sémantique, première version.

3.3 Modèles d'intégration de données

Avant le développement de l'internet (hautement hétérogène et réparti), diverses approches d'intégrations de données ont été proposées afin de regrouper de l'information venant de plusieurs sources de données dans un ensemble homogène pour faire de l'analyse de données. Elles se divisent en 3 grandes approches : centralisée, fédérée et partagée. Ces approches visent à définir une méthodologie de partage d'information tant au niveau des modèles de données structurant l'information, que de leur vocabulaire. Ces approches sont utilisées pour le partage de données inter-systèmes d'information mais sont aussi utiliseés à un niveau local (hôpital, entreprise, etc.). Par exemple, les besoins au niveau d'un système d'information hospitalier ne sont pas les mêmes que ceux que peuvent avoir des chercheurs qui veulent avoir accès aux informations de n hôpitaux depuis une vue unique. Les opérations locales doivent être supportées par un système d'information qui s'adapte au métier, qui soit flexible, robuste, et où l'information doit être sécurisée et non volatile. L'alignement d'une multitude de sources de données, par exemple pour de l'analyse décisionnelle, obéit à des critères de partage de sens de l'information, de possibilité d'agrégation de l'information ou bien de représentation de l'information pour analyse (la sécurité est bien entendu une propriété partagée). Les approches que nous présenterons dans cette section sont applicables au niveau local et global, elles n'ont cependant pas les

mêmes propriétés suivant les domaines d'application.

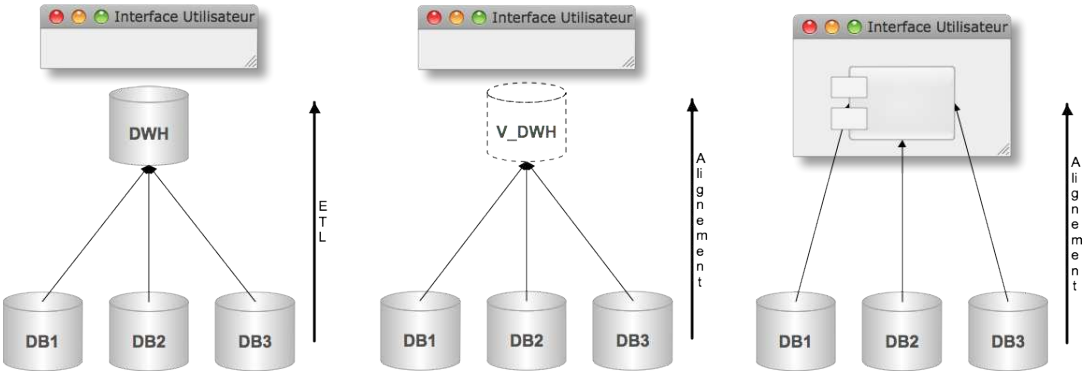


FIGURE 3.2 – 3 approches d'intégration de données : 1. L'approche centralisée où les données sont stockées de manière persistante dans un entrepôt de données (DWH), 2. L'approche fédérée où l'entrepôt de données est non persistant (V_DWH) et 3. L'approche "mash up" ou de partage de modèles où les données sont utilisées à la volée par l'application.

On distinguera trois approches dans l'intégration de données multi-sources (Figure 3.2). Dans une approche classique, dite centralisée, les données sont matérialisées dans un entrepôt de données central et intégrées via des processus ETL [3] . Dans l'approche fédérée ou de médiation, plus récente, les données sont matérialisées mais à leur source. L'entrepôt est alors « virtuel », et le processus d'intégration est implicitement effectué par rapprochement de modèles de données identiques (fédération) ou par fusion de modèles de sources différentes (médiation). Enfin, l'approche partagée propose de relier les différentes sources entre elles de manière indépendante les unes des autres, le plus souvent au sein de l'application même grâce à l'utilisation de services web. Le coût de mise en oeuvre de chaque approche est différent. L'approche centralisée où un modèle d'information est créé et où les données sont intégrées (et souvent vérifiées et normalisées) dans un entrepôt unique est coûteux mais reste la plus simple à mettre en oeuvre malgré les problèmes de confidentialité des données. L'approche fédérée permet de pallier la problématique de persistance des données (voir la section suivante) mais ne résout pas le problème d'alignement des modèles et des vocabulaires. L'approche médiée permet de s'affranchir de la problématique du modèle unique mais pose la problématique d'alignement de modèles. Enfin, l'approche partagée, plus adaptée à l'internet, propose d'utiliser l'information telle qu'elle, mais posera des problématiques d'alignement et d'identification

3. Extraction, Transformation et Chargement (Load) des données

des instances. Dans ces quatre approches, les problématiques restent quasiment les mêmes ; il faut aligner des structures d'information (modèles) et des vocabulaires afin de pouvoir interroger conjointement plusieurs bases de données.

3.3.1 Intégration centralisée et persistante

L'approche la plus répandue pour intégrer des données reste l'approche centralisée. On intègre dans un même modèle des données issues de sources hétérogènes. Cette intégration se fait de manière manuelle et l'interprétation des données à intégrer est confiée à l'homme. Cette approche est populaire car elle garantie des performances d'interrogation élevées et elle garantie une intégration contrôlée des données tant au niveau du vocabulaire que de la sémantique. Lors de cette intégration de données, différentes opérations de qualité de données seront mises en oeuvre (voir Chapitre 2). Un autre avantage de cette approche est la séparation de l'entrepôt de données vis à vis du système opérationnel. Il n'est pas toujours possible de faire des requêtes générant des résultats volumineux et coûteux pour le moteur du SGBD directement sur un système opérationnel. Par exemple, dans un hôpital, le système de gestion du dossier patient est hautement sécurisé et déjà très sollicité par les tâches opérationnelles. Rajouter un système d'analyse de données sur une population de patients poserait des problématiques opérationnelles lourdes. On le verra plus tard, la mise en oeuvre d'un entrepôt de données couplé à une approche de médiation sémantique n'est pas exclue pour les raisons citées. La procédure de mise en oeuvre classique d'un entrepôt de données est la suivante (après définition des besoins métiers, etc.) :

- Analyse des données et des modèles de données sources
- Modélisation conceptuelle et le plus souvent dimensionnelle du domaine d'étude
- Spécialisation du modèle conceptuel en un modèle logique et physique de base de données (le plus souvent relationnel)
- Chargement des données via des procédures ETL incluant la mise en oeuvre de procédures qualité, de standardisation et d'adaptation des données (dédoublement, gestion des conflits, etc.)
- Mise en oeuvre du moteur OLAP [4] (pour la navigation en ligne dans les données en temps réel) (option)

L'analyse des bases de données sources permet de construire, le plus souvent manuellement, le modèle de données de l'entrepôt cible en fonction des besoins d'analyses. Diverses recherches afin de semi-automatiser ou d'automatiser le processus de

4. OnLine Analytical Processing

génération du modèle d'information de l'entrepôt de données ont été proposées.

Outre l'aspect d'alignement du modèle de données, il faut des méthodes d'alignement des données. L'alignement de celles-ci vers l'entrepôt de données se fait suivant différentes approches. D'abord manuellement. Dans les années 90, [Deen 1987] propose des premiers travaux d'intégration de données manuelles, essentiellement au niveau de la syntaxe des données. Les approches d'intégrations manuelles visent à fournir des outils et des langages de manipulation de données pour l'intégrateur. Ces outils sont très populaires et sont généralement connus sous l'acronyme ETL (extract, transform and load). Des approches semi-automatisées et automatisées existent, mais sont généralement très spécifiques. La communauté de recherche des entrepôts de données s'interrogent sur la possibilité de générer des entrepôts de données à la demande, en fonction des besoins d'analyse. Les coûts de transfert de l'information restent élevés, mais si l'entrepôt de données construit de façon dynamique (on parle d'entrepôt de données actif) est suffisamment restreint et spécifique, cette solution est intéressante [Thalhammer 2001]. L'approche proposée par Thalhammer pour construire des entrepôts de données actifs (donc des espaces d'analyse) montre la complexité de modélisation du processus de décision et de l'applicabilité restreinte de celle-ci dans le cas d'une approche automatisée. Nous verrons plus tard comment dans ce cas précis, l'usage d'ontologie semble adapté au processus décisionnel et à la constitution de "cubes de données" à la volée en fonction de celle-ci. Une limite de l'approche centralisée est la complexité de la gestion du changement du schéma et des données. Elle reste lourde à gérer bien que des approches aient été proposées. Il est d'ailleurs courant de résoudre le problème en rechargeant entièrement les données à chaque mise à jour de l'entrepôt de données, ce qui est très lourd en temps de gestion. Le projet WHIPS développé à l'université de Stanford vise à la mise en oeuvre d'une architecture d'entrepôt de données qui est capable de s'adapter avec le temps grâce à un système de gestion de vues. En effet, lorsqu'un modèle de données est mis en oeuvre, la modification de celui-ci entraîne invariablement un nouveau chargement complet de l'entrepôt de données ; c'est une tâche très lourde. La figure 3.3 présente l'architecture de maintenance de l'entrepôt issu de WHIPS.

Chaque source de données est associée à une vue. Ces vues sont associées à un système de gestion de vues qui permet de mettre en oeuvre une synchronisation des vues entre les sources en fonction de leur évolution. L'intégrateur de vues s'occupe ensuite de gérer l'alignement entre les sources.

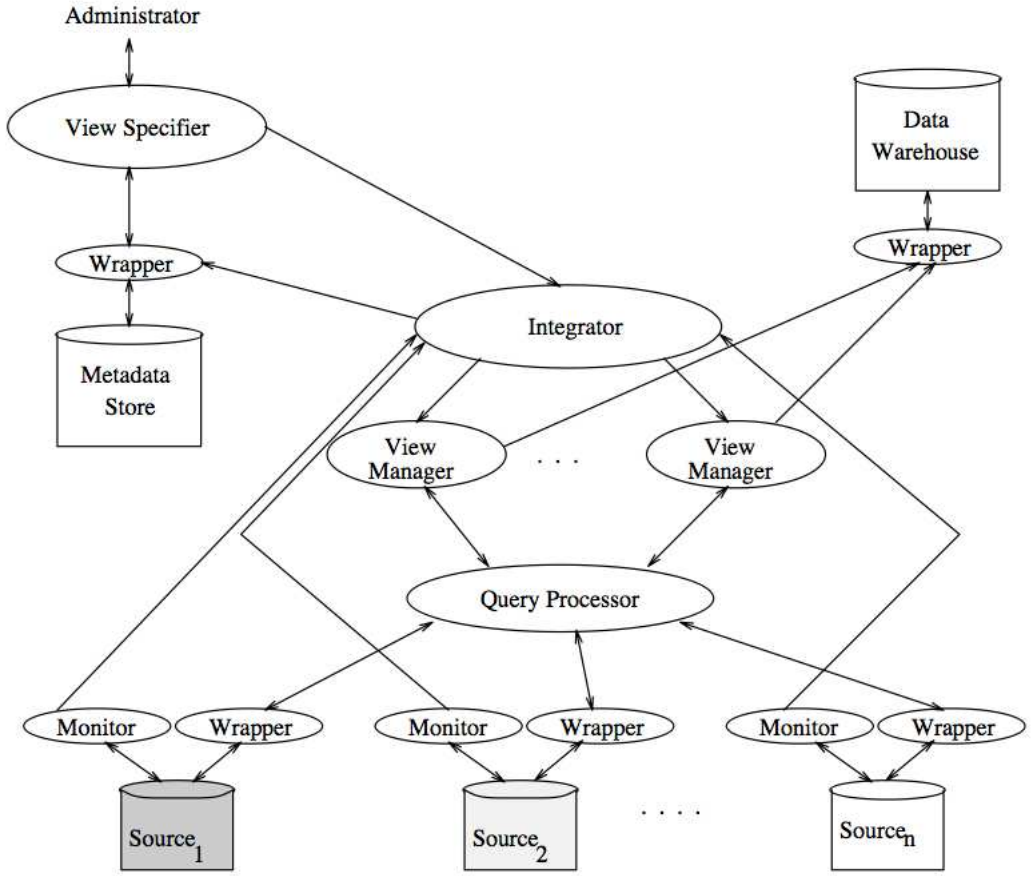


FIGURE 3.3 – L'architecture du projet WHIPS.

3.3.2 Intégration centralisée ou décentralisée non persistante

Un entrepôt fédéré est selon [Sheth 1990] une collection de bases de données collaboratives, autonomes et hétérogènes. Ce type d'approche propose d'aligner les modèles d'information sources vers un modèle pivot commun mais sans stocker les données dans un entrepôt de données de manière persistante. Les processus d'intégration sont executés à la volée sur les bases sources une fois qu'elles ont été définies. Nous distinguerons deux types d'architectures dans cette approche :

- Architecture fédérée : les bases de données sources ont le même schéma et le schéma global est identique aux sources. Le système de requête central relaie les requêtes vers chaque site et agrège les résultats pour l'affichage.
- Architecture médiée : les bases de données sources n'ont pas le même schéma et le schéma global est soit une agrégation des modèles sources, soit un schéma

unique et différent qui sera lié aux schémas source grâce à des règles d'alignement par exemple. La requête est envoyée et traitée par le médiateur qui transforme la requête pour l'adapter aux schémas source. Le résultat est agrégé par le médiateur.

L'alignement d'ontologies (ou de modèles de données) dans le cadre d'architecture médiée se fait suivant 3 approches :

- GAV (global as view) : une ontologie globale est utilisée comme source de médiation, chaque fournisseur de données lie ses données à cette représentation pivot.
- LAV (local as view) : chaque source est définie à l'aide d'une ontologie locale (ou ontologie de données), l'alignement entre ces ontologies locales peut être difficile.
- GLAV (global as local as view) : une ontologie globale est construite à partir des ontologies locales de données.

Les types d'architecture ont le même but que pour l'approche centralisée persistante, c'est à dire de proposer une vue unique (agrégée) des données pour l'interrogation, ceci facilitant l'interrogation des données par des utilisateurs ou des applications tierces. La principale différence entre ces approches réside dans l'aspect de stockage persistant où non des données. L'avantage du système d'alignement de modèles via la fédération ou la médiation est donc la fraîcheur des données. En effet, les données restent dans les sources et ne sont pas centralisées. Elles sont donc récupérées au moment voulu, de manière dynamique. La problématique citée de ces approches est la performance. Il est généralement conseillé de bien limiter les requêtes faites en pré-traitement comme dans le projet PICSEL qui a donné lieu à un projet de construction de médiateurs semi-automatisée grâce à l'utilisation d'ontologies pour formaliser les sources de données et le domaine [Reynaud 2002].

3.3.3 Intégration à la volée ou "mashup"

Les applications composites (ou "mashups") sont des applications qui combinent des ressources de données et/ou de services pour créer une nouvelle application. Dans le cas de l'intégration de données, le mashup est le fait d'interroger des sources d'information à la volée au sein d'une même application. La vue unique sur les données n'est donc pas nécessaire bien qu'il faille gérer l'intégration des données dans l'application, mais sans l'utilisation nécessaire d'un modèle pivot. Le mashup est en quelque sorte une médiation de données où le médiateur est inclus dans l'application cible ou dans un service tiers qui sera appelé au moment de l'intégration de données. Cette approche est co-notée d'un aspect service, ce qui la différencie de

l'approche de médiation. L'intégration à la volée est certes séduisante, mais cette approche reste complexe à mettre en oeuvre notamment à cause du manque d'outils pour gérer les alignements entre sources de données. Par exemple, construire un mashup de données autour d'un sujet sur l'internet reste complexe si les deux sources ne se sont pas alignées au niveau sémantique.
GoogleMaps [5] est l'API la plus utilisée dans des applications web mashup aujourd'hui. Par exemple, une utilisation de googlemaps est faite avec youtube afin d'avoir une vue globale et mise à jour en temps réel des vidéo ayant pour sujet le conflit à Gaza (http://www.mibazaar.com/gazaconflict.html). Les mashups sont de plus en plus utilisés sur le web, notamment grâce à l'avènement des micro-formats, du web social, et plus récemment du web sémantique.

3.4 Le Web Sémantique

D'une manière naturelle, le web est devenu la plateforme d'échange d'information par excellence. Cet échange se fait massivement par l'intermédiaire de la publication de documents (pages html) reliés entre elles grâce au système d'hyperlien. La structure même du réseau (l'Internet) permet facilement l'interconnexion de milliards de pages web et des outils de recherche dans ces pages via des systèmes d'indexation et de classement (google) ont été rapidement développés face à la quantité d'information présente. Il se pose aujourd'hui un problème de masse de données et de recherche dans cette masse. Deux idées sont proposées par le fondateur de la couche logique d'Internet (le world wide web) [Berners-Lee 2001] : 1) il ne faut plus penser en terme de documents, mais en terme de données (unité la plus petite possible) et 2) il faut que ces données aient un sens clair et qu'elles soient identifiables. Ces deux principes, simples, sont les fondements du web de données (LinkedData) définis par Tim Berners-Lee en 1999 lors de la publication de la première version de RDF [6]. L'apparente simplicité du modèle de données RDF donne au web sémantique un essor dans toutes les communautés, mais en particulier dans la communauté des sciences de la vie [7]. En effet, le caractère hautement hétérogène et réparti de l'information sur l'Internet est très similaire à l'information de santé. Il en est de même pour le besoin d'organisation de cette masse d'information. La figure 3.4 montre l'état du chantier web sémantique actuel. Cette représentation met en avant la composante complexe des standards et des langages du web sémantique. Les outils commencent cependant à émerger et leur utilité ainsi que leur validité dans le domaine de la santé

5. http://www.googlemaps.com
6. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/
7. Le Semantic Web Health Care and Life Science Group : http://www.w3.org/blog/hcls

doivent encore être prouvés. Il est important d'effectuer une lecture de ce "gâteau" dans le cadre de cette thèse, particulièrement pour mieux caractériser les frontières entre les niveaux. Nous lisons la pile du web sémantique du niveau le plus bas, vers le niveau le plus haut :

- Technique - La plateforme logique du world wide web : le processus d'identification des objets(URI), le protocole de transport (HTTP) et les services liés à la sécurité (AUTH).
- Syntaxique - Les formats d'encodage des données (XML, TURTLE, RDFa) et le modèle de données RDF.
- Sémantique - Un modèle et un formalisme (langage) permettant de modéliser et de capturer la connaissance d'un domaine avec un niveau d'expressivité différent suivant le langage choisi (OWL-DL, OWL-Full, SKOS, RDFS, DAML+OIL).
- Sémantique - Un langage de règles permettant d'exprimer des relations entre des concepts d'un domaine qui sont définis par exemple dans une ontologie (OWL). A la différence d'une ontologie, les règles représentent des relations conditionnelles entre des concepts.
- Logique - Transversalement à la connaissance stockée dans des ontologies ou bien des règles, la logique est utilisée pour générer de nouvelles connaissances ou pour interpréter les connaissances des niveaux inférieurs de la pile. Plusieurs types de logiques existent, la logique de description et la logique de premier ordre en sont deux exemples. La logique peut utiliser le raisonnement pour induire de nouvelles règles. On distinguera plusieurs raisonnements : inductif, déductif et abductif.
- Preuve - Les preuves représentent les étapes du raisonnement logique en fournissant des explications de celui-ci.
- Vérité - Nous admettons ici que les preuves représentent les axiomes du domaine de discours. Les axiomes sont des règles qui ont un niveau de vérité suffisant dans un domaine donné pour être utilisées afin de raisonner sur ce domaine. Par exemple, le fait qu'un 'médecin' soit un homme dans une ontologie décrivant la clinique est une connaissance implicite qui est admise comme vraie et qui est inutile à démontrer.

Nous observons aussi dans la figure 3.4 que le web de données est un sous-ensemble du gâteau. Le web de données est aux données ce que le web est aux pages web. Le web de données est un projet qui vise à partager et inter-lier des données sur le web. Pour ce faire, le web de données propose d'utiliser le web comme mécanisme

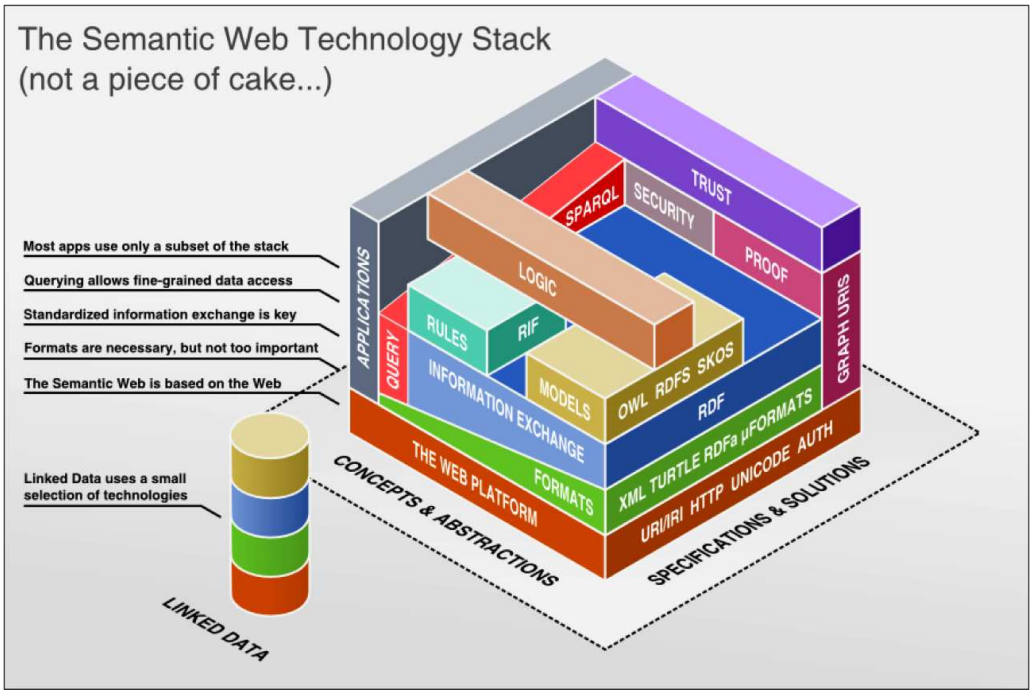


FIGURE 3.4 – La pile du web sémantique actuelle. Crédits : http ://bnode.org/blog/2009/07/08/the-semantic-web-not-a-piece-of-cake

de transport (http) et d'identification (URI) des données modélisées grâce à RDF. Ces données sont annotées à des ressources sémantiques qui seront elles liées à d'autres ressources sémantiques (ontologies). La figure 3.5 représente une vue logique du mécanisme de liaison entre les données sur le web de données : les bases de données sont liées à des concepts (things) qui sont eux-mêmes reliées entre eux. Des applications peuvent ensuite utiliser les données.

La figure 3.6 (en fin de chapitre) présente l'état du nuage "linked data" aujourd'hui du point de vue des ressources actuellement liées. Le projet est en expansion et la recherche autour de ces sujets s'organise. Les outils et méthodes de partage d'information sur le Linked Data ont la capacité de gérer les vocabulaires, les terminologies et les ontologies en plus des données. Standardiser le contenu et la structure de partage de l'information est un moyen d'aider à l'intéropérabilité.

3.4.1 Langages de représentation

Le W3C a présenté OWL [8] en 2004. Ce langage de représentation de l'information est adapté pour formaliser des connaissances que la machine peut traiter. Les on-

8. Ontology Web Language : http://www.w3.org/TR/owl-features/

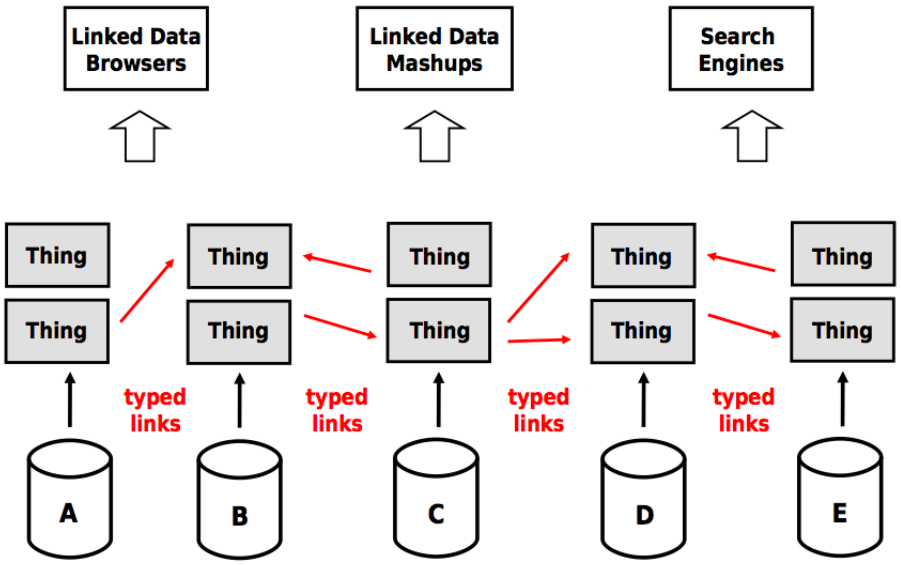


FIGURE 3.5 – Une vue logique du mécanisme de liaison des données par leur sémantique (ou métadonnées) sur le web de données.

tologies sont généralement représentées sous forme de graphes où des concepts sont reliés par des relations (voir Chapitre 2). OWL est un aboutissement de l'évolution des langages de représentation de l'information sur le web.

- XML fournit la syntaxe pour des documents structurés mais n'impose pas de sémantique sur ces documents,
- XML Schema est un langage pour restreindre la structure des documents XML et introduit la notion de type de données,
- RDF est un modèle de données pour des objets et des relations entre ces objets. RDF peut être représenté en syntaxe XML ou N3.
- RDF Schema (ou RDF-S) est un vocabulaire permettant de décrire les propriétés et les classes des ressources RDF avec une sémantique permettant de décrire des notions de hiérarchies (généralisation - spécialisation), par exemple.
- OWL rajoute du vocabulaire pour décrire les classes et les propriétés, parmi lesquelles on retrouve les relations entre les classes (disjonction), la cardinalité (exactement une), l'égalité, les caractéristiques des propriétés (par exemple la symétrie), etc.

La richesse du langage OWL est décliné en trois niveaux. OWL-Lite, est plus adapté à la représentation des classifications avec des contraintes simples (0 ou 1). Historiquement, OWL-Lite a été proposé pour répondre à des besoins plus simples et pour faciliter la mise en oeuvre d'outils autour du langage. OWL-DL (Description

Logics) est adapté aux personnes désirant un haut niveau d'expressivité, tout en restant calculable. Enfin, OWL-Full est la représentation la plus libre et la plus expressive. Par exemple, une classe en OWL-Full peut être traitée comme un concept où comme un individu. D'un point de vue logique, pour que le langage soit compréhensible et inférable par la machine, le langage doit être restreint en terme de vocabulaire, afin, par exemple, que les propriétés qui relient deux concepts aient un sens logique (par exemple la transitivité, ou l'équivalence). C'est pourquoi une partie de la communauté pense que le langage doit être structuré, et que cette structure doit être interprétable par un raisonneur. Une autre partie de la communauté pense que l'information est par nature liée à un contexte qui même lorsqu'une ontologie est disponible, est hautement contextuelle en fonction de l'usage que l'on veut faire de l'information. Dans cette communauté, les raisonneurs développés tentent de mettre en oeuvre des logiques pour empêcher les phénomènes de boucle sur les raisonnements, ceci afin de permettre l'utilisation d'ontologies de type OWL-Full.

3.4.2 Langages de règles

Les langages à bases de règles sont utilisés pour stocker des relations propositionnelles ou non entre des concepts. Ces langages sont généralement liés à leur capacité d'exprimer des raisonnements issus des logiques. Pour W.G. Leibniz [Leibniz 1886], avant de savoir si ce que l'on pense est vrai, il faut s'assurer de la logique de son raisonnement. C'est pourquoi, en logique formelle, il énonce le principe de contradiction où il dit qu'une proposition ne peut être ni vraie ni fausse en même temps car dans ce cas, la proposition serait indécidable. Sans se soucier de la vérité de la proposition elle-même. Cet axiome explique la problématique de mise en oeuvre de règles et définit l'espace dans lequel celles-ci peuvent s'appliquer. La vérité de la proposition (comme nous l'avons vu dans la section précédente), est cependant importante en logique. Le syllogisme suivant : "les éclairs au café sont des gâteaux, les gâteaux sont vendus chez les pâtissiers, donc les éclairs au café sont vendus chez les pâtissiers." est vrai. Par contre, le paralogisme : "les éclairs aux cafés sont des choses délicieuses, des choses délicieuses sont vendues chez les glaciers, donc les éclairs au café sont vendus chez les glaciers." est faux. Toutes les prémisses sont pourtant vraies. Nous voyons donc dans cet exemple que la véracité d'un raisonnement ne réside pas seulement dans sa forme (logique formelle), elle dépend aussi de la signification des mots qui y figurent. C'est une des raisons qui pousse la communauté du web sémantique à utiliser des systèmes de règles formelles couplées à des espaces de connaissances (ontologies).

3.4.2.1 SWRL / Semantic Web Rule Language

SWRL, proposé au W3C en 2004, est un langage à base de règles qui utilise les langages ontologiques OWL-Lite/OWL-DL et RuleML. Cette caractéristique le rendant indécidable, SWRL DL-safe rules est une restriction de SWRL conçue pour conserver la décidibilité [9]. Les règles SWRL sont supportées par de nombreux moteurs d'inférence : Bossam, Hoolet, KAON2, Pellet, RacerPro, R2ML.

Voici un exemple de règle SWRL :

```
aParent(?x1,?x2) ^ aFrère(?x2,?x3) => aOncle(?x1,?x3)
```

Cette règle, implique que si Pierre a Margot comme parent, et que Margot a Philippe comme frère, alors Pierre a Philippe comme oncle. Par contre, une règle de la forme :

```
Etudiant(?x1) => Personne(?x1)
```

Cette règle duplique la propriété owl :sameAs qui serait directement interprétable par un raisonneur. Il n'est donc pas généralement préférable de faire ce type de règle . SWRL implémente un certain nombre de 'built-ins' qui permettent la comparaison(swrlb :equal, swrlb :lessThan, etc.), des expressions mathématiques (swrlb :add, swrlb :round, etc.) ou bien la gestion des chaînes de caractère (swrlb :stringConcat, swrlb :contains, etc.). Les définitions de ces built-ins sont accessibles à l'espace de noms http ://www.w3.org/2003/11/swrlb.

3.4.2.2 Turtle N3

Turtle N3 (ou notation 3) n'est pas, à proprement parler, simplement un langage de règles. C'est d'abord une sérialisation de RDF, tout comme l'est XML, ou plutôt RDF/XML. N3 est une manière plus lisible d'écrire des triplets RDF (donc RDF/N3)., comme le montre l'exemple suivant :

```
:Mary :son :Frank, a :Male;
      :son :Bob, a :Male;
      :son :Sam, a :Male.
```

Pour le cas des règles, un framework N3Logic, basé sur la notation 3 est proposé. Voici dans cet autre exemple une règle qui permet d'axiomatiser le fait que tous les fils d'âge inférieur à 15 ans sont des garçons.

9. La décidibilité est la capacité d'un langage à contenir des boucles infinies de traitement par un automate

```
{ ?x :fils ?y.
       ?y :age ?z.
       ?z math:lessThan 15. }
 =>
       { ?y a :Garçon. }
```

Nous remarquons dans cet exemple que la notation 3 permet l'utilisation de { } afin d'imbriquer des blocs de propositions. Cela donne à N3 la possibilité de mettre en oeuvre des structures hiérarchiques, et donc des relations n-aires. L'exemple suivant nous montre comment il est possible d'écrire une règle n3 disant que tous les parents de mammifères sont des mammifères.

```
{ ?x a : mammifères. }
  =>
{ ?x :parent [ a : mammifères ]. }
```

Comme SWRL, le langage de règles N3 propose un ensemble de built-ins permettant d'effectuer des opérations sur des concepts.

3.4.3 SPARQL - Standard Protocol and RDF Query Language

Le langage de requêtes SPARQL est un langage de type SQL permettant d'interroger des données modélisées en RDF. Pour exprimer des graphes RDF dans la partie conditionnelle de la requête, la syntaxe N3 est utilisée. SPARQL est capable d'exprimer une requête et de retourner des résultats en utilisant la formalisation d'une ontologie. SPARQL est vu par le W3C comme le langage de requête et le protocole de transport des données sur le Web de Données.

Voici un exemple d'utilisation de la clause SELECT en SPARQL :

```
PREFIX foaf:   http://xmlns.com/foaf/0.1/
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name .
        ?x foaf:mbox ?mbox . }
```

La première ligne définit un espace de nom (foaf : Friend Of A Friend ontology), les deux lignes du WHERE utilisent le préfixe $foaf$ pour exprimer le graphe RDF à comparer. Il est possible d'utiliser des contraintes (FILTER) pour restreindre la sélection de données RDF à un sous-ensemble remplissant une contrainte. Par exemple :

```
PREFIX foaf:   http://xmlns.com/foaf/0.1/
SELECT ?name ?mbox
WHERE { ?x foaf:name ?name .
        ?x foaf:mbox ?mbox . }
FILTER regex(?mbox, "inserm.fr")
```

La même requête est possible avec un filtre qui utilise une clause de type expression régulière pour récupérer seulement les informations relatives à une personne qui a une adresse mail en "@inserm.fr".

La partie WHERE d'une requête SPARQL peut être optionnelle (OPTIONAL). Dans ce cas, la partie optionnelle est évaluée lorsque le triplet comparé est présent, mais la comparaison n'échoue pas si la partie testée est absente. Il est aussi possible d'unir plusieurs graphes (UNION) et de les comparer. À ce jour, d'autres opérateurs sont disponibles dans SPARQL 1.0 que nous retrouvons dans SQL, comme ORDER BY, DISTINCT ou bien LIMIT. L'opérateur DESCRIBE permet de retourner de l'information à propos de la ressource RDF. Il est aussi possible de construire des graphes nommés grâce à la clause CONSTRUCT dans laquelle on va lier les concepts créés dans le CONSTRUCT avec les concepts du WHERE grâce aux variables.
SPARQL 1.1 est proposé en version brouillon pour le moment. Cette évolution de SPARQL vise entre autres à proposer des opérateurs GROUP BY (qui permettront de réduire la charge des requêtes faites sur les SPARQL endpoint). Il y sera aussi implémenté un mécanisme de mise à jour des données (UPDATE) et d'effacement (DELETE). Enfin, cette version permet le filtrage sur une négation. Cette révision propose aussi la mise en oeuvre d'opérateurs permettant certains types d'inférence directement en SPARQL.

3.4.4 Logiques et Raisonneurs

La formalisation de la sémantique d'un domaine, permet l'utilisation de logiques pour inférer ou déduire des faits ou des concepts. OWL est un langage sémantisé (muni d'une sémantique formelle) et un langage de raisonnement (un langage sémantisé ayant des méthodes de raisonnement intrinsèques et mathématiquement définies). Un langage sémantisé permet de définir une fonction d'interprétation entre un élément (syntaxe) et sa signification (sémantique).

Pour établir la sémantique formelle d'un langage L on choisit un langage logique ou mathématique S (sémantique) qui fait autorité, comme par exemple la théorie des ensembles, la logique du premier ordre, les distributions de probabilité, etc. On établit ensuite une correspondance fonctionnelle des éléments atomiques de

la syntaxe de la théorie choisie (une interprétation). Eventuellement, on choisit si le langage possède certaines propriétés en rapport avec sa sémantique comme par exemple la non-contradiction (une expression ne doit pas conduire à la déduction de son contraire), la monotonie (une nouvelle connaissance apprise ne réduit pas l'ensemble des connaissances déduites), etc.

Les logiques sont basées sur des théories. La logique propositionnelle, la logique des prédicats ou la logique du premier ordre sont, par exemple, basées sur la théorie des ensembles. La logique floue est basée sur la théorie des ensembles flous. etc. Chaque logique à donc sa propre sémantique. OWL se décompose en plusieurs langages différemment expressifs qui se basent sur des logiques différentes :

- OWL-Full peut représenter les connaissances sans restrictions mais n'offre pas la possibilité d'un raisonnement automatique (à part dans un cas où le raisonneur permet de détecter d'éventuelles boucles de raisonnement et où il est possible de préciser la sémantique à utiliser)
- OWL-DL (description logics) permet de représenter les connaissances avec quelques restrictions mais reste décidable dans un temps exponentiel de la taille de la donnée
- puis OWL-Lite, OWL-EL, OWL-RL, OWL-QL sont décidables avec des limitations adaptés à certains usages (ontologies volumineuses, bases de données, etc.).

Les raisonneurs sont des outils logiciels capables d'effectuer des raisonnements (d'appliquer des mécanismes logiques) pour déduire des faits à partir d'une base de connaissances. Il existe 2 grandes familles de raisonneurs. Un raisonneur à chainage avant démarre avec des faits et utilise des règles d'inférence pour générer de nouveaux faits. Ces nouveaux faits sont rajoutés à la base de faits qui peut générer de nouvelles règles, puis le procesus continue jusqu'au moment où il n'y a plus rien à inférer. Un raisonneur à chainage arrière démarre avec un jeu d'hypothèses et cherche dans les règles d'inférence les règles qui ont les hypothèses dans leur objet (sujet => objet). Par exemple :

- Prenons 100 employés dans une entreprise, nous avons donc 100 faits du type :Pierre :estEmployéDe :entreprise
- Il y a aussi 100 matériels informatiques différents dans cette entreprise, donc 100 autres faits du type :entreprise :aMatériel :imprimanteM1
- Une règle de contrôle d'accès simple serait de dire qu'une personne employée de l'entreprise peut utiliser le matériel de celle-ci :

```
{?personne :estEmployéDe ?entreprise. ?entreprise :aMatériel ?matériel}
  =>
```

```
{?personne :peutUtiliser ?matériel}
```

- Un raisonneur de type "forward-chaining" va générer 10000 nouveaux faits et les rajouter à la base de faits. La vérification qu'un employé peut utiliser un matériel se réduit à un problème d'interrogation de la base.
- Un "backward-chaining" raisonneur fera du raisonnement à la demande, si :Pierre utilise :imprimanteM1, le raisonneur recevra :Pierre :peutUtiliser :imprimanteM1 et l'hypothèse sera évaluée comme vraie ou fausse.

La plupart des raisonneurs actuels sont à chainage avant. Euler [10] est un générateur de preuves à chainage arrière. Dans le cadre du projet DebugIT, Euler sera notre choix de raisonneur pour assurer l'interopérabilité entre les différentes sources de données et l'ontologie de domaine.

3.5 Conclusion

Nous avons vu dans ce chapitre différentes approches utilisées dans la littérature afin d'assurer l'intégration de données pour l'analyse. Puis, nous avons exploré des méthodes permettant d'accéder à l'interopérabilité sémantique entre systèmes, sans intégrer les données dans une vue particulière à proprement parlé. Le web sémantique, de part sa structure ouverte (grille), propose aujourd'hui des méthodes qui peuvent aider à intégrer des données à la volée. Il propose même d'aider à construire une interopérabilité sémantique entre différents systèmes d'information. Le web sémantique ne résout cependant pas tous les problèmes. Nous aimerions insister sur le terme sémantique. La sémantique est l'étude des signifiés, autrement dit, de la signification des mots, des rapports entre eux et de leur contexte d'utilisation. Autrement dit le web sémantique doit apporter des solutions afin d'identifier le sens (dans son contexte) des signes (mots, images, etc.). Les ontologies sont des ressources qui permettent d'aider à ce processus d'identification du sens d'un concept. Il reste cependant difficile, même grâce à des ontologies, de modéliser le contexte d'interprétation d'un concept. Le web de données vise à interconnecter des données qui ont le même sens entre elles. Le web de données n'a pas besoin à proprement parler d'ontologies pour faire cela, il peut le faire sur un mécanisme d'interconnexion basé sur des métadonnées et non nécessairement sur des ontologies formelles. L'apport de l'ontologie pour interconnecter des données n'est donc pas si clair aujourd'hui dans la communauté du web sémantique. Dans tous les cas, il est important de dissocier les problèmes relatifs à la sémantique (au sens formel) et les problèmes relatifs à l'intégration de données à grande échelle (linked data). La sémantique

10. http://www.agfa.com/w3c/euler/

peut aider à l'intégration de données, mais n'est pas forcément nécessaire. L'étude de la sémantique est donc pour nous déconnectée de la problématique d'intégration de données. Cependant, comme la communauté du web sémantique (au sens formel), nous pensons que si les données sont sémantiquement identifiées de manière formelle, l'intégration de données est plus facile. Il n'en reste pas moins, et nous le verrons dans la suite de cette thèse, que la problématique de l'annotation de données et de l'interconnexion données-connaissances est une problématique de recherche à part entière qui nécessite des processus de raisonnement que nous venons de vous présenter succinctement. Dans le chapitre suivant, nous allons aborder les notions précédemment traitées dans les chapitres 2 et 3 mais dans le contexte de l'informatique médicale.

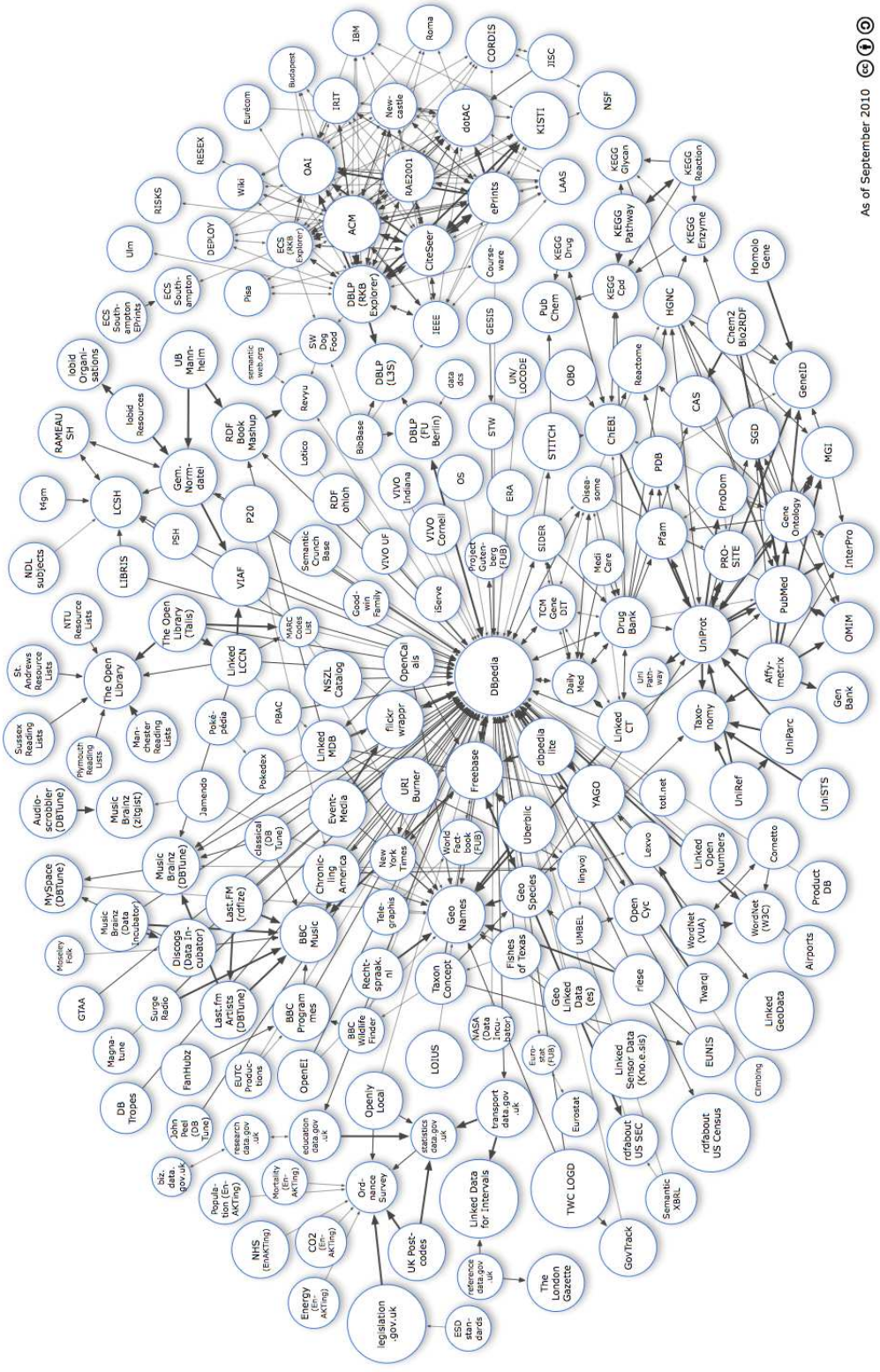


FIGURE 3.6 – Linked Data, Statut du Data Cloud en Septembre 2010.

Chapitre 4

Standards en santé pour partager l'information

"L'être humain n'a aucun standard de qualité, hormis son besoin d'appartenance." - Bernard Arcand

Sommaire

4.1 Introduction 72
4.1.1 Modèles : Health Level Seven (HL7) 72
4.1.2 Modèles : openEHR 74
4.1.2.1 Le modèle d'information d'openEHR 75
4.1.2.2 Les archétypes openEHR 75
4.1.2.3 Différences et complémentarités 77
4.1.3 Terminologies : NEWT / ATC / ICD10 78
4.1.4 Ontologie : SNOMED CT et UMLS 79
4.2 Interopérabilité Sémantique en Santé 80
4.2.1 Aggrégation de modèles, approche entrepôt de données . . . 82
4.2.2 Alignement de modèles : fédération et médiation de données . 83
4.2.2.1 Intégration à la volée et "mashup" 84
4.3 Analyse Comparative 85
4.4 Synthèse et Discussion 86

Le standard reste une réponse valable pour aider à l'interopérabilité. En effet, si tout le monde utilise le même standard, nous sommes à peu près sûrs du sens du concept utilisé, puisqu'il a été défini dans un standard et utilisé par l'homme. Nous savons cependant aussi, que le standard peut avoir tendance à appauvrir le contenu informationnel qu'il code. Pour être consensuel, le standard doit s'adapter à tout le monde, et dans le cas de l'information biomédicale, c'est chose quasiment impossible. Si on essaye de coder de manière explicite l'information dans un standard,

alors il devient si complexe qu'il est lourd à mettre en oeuvre par l'homme et à traiter par la machine. Cependant, nous savons aussi que mettre en oeuvre un standard demande la mise en place d'efforts humains colossaux qui aboutissent très souvent à une amélioration de la qualité des données générées, même si le codage appauvrit l'expressivité de ces données. Il reste donc important d'utiliser ces standards quand on le peut et de tenter de les faire évoluer.

4.1 Introduction

Il existe en informatique de nombreux standards pour permettre aux systèmes de communiquer. Le domaine bancaire définit par exemple, des nomenclatures de produits financiers au niveau international. Il définit aussi un système de codage de comptes partagé, et des messages codés permettant la communication entre les différents établissements bancaires. En santé, le besoin de partage d'information se faisant grandissant, des groupes se penchent aujourd'hui sur la mise en place de standards afin de coder et de décrire les données, ainsi que des protocoles pour les échanger. Nous présentons dans cette section diverses initiatives de standardisation de l'information en santé. Qu'elles soient nationales ou internationales, ces initiatives sont comparables à ce qui a été proposé dans d'autres domaines, à la différence qu'il est difficile d'imposer un standard ou une nomenclature en santé puisque le praticien n'a pas directement "besoin" des systèmes d'information de santé pour soigner. Nous verrons cependant dans la suite de cette thèse que le standard améliore la qualité de l'information enregistrée mais qu'il n'est pas absolument nécessaire pour créer de l'interopérabilité.

4.1.1 Modèles : Health Level Seven (HL7)

HL7 est une organisation (originairement américaine qui devient internationale) accréditée ANSI et ISO qui vise à proposer des standards liés aux échanges d'informations médicales entre applications cliniques. Des groupes de travail définissent, par domaine médical, des modélisations conceptuelles formalisées en UML. Ces modèles UML sont alors partagés dans le ballot d'HL7.

HL7 définit la structure et le rôle des messages entre applications. HL7 est un standard qui a été largement adopté pour assurer le transport de messages inter-applicatifs dans sa version 2. Malgré la standardisation des messages, il demeure difficile d'assurer une interopérabilité complète entre différents instituts de soins. Tout d'abord parcequ'il n'y a pas de lien clair entre les différents domaines standar-

disés dans HL7 v2 et aussi parceque la standardisation du contenu des messages n'est pas gérée. C'est pour pallier ces problématiques, que HL7 version 3 a été introduit. HL7 version 3 est d'abord un standard d'échange de messages, mais est en train de devenir un standard pour aider à la modélisation des systèmes d'informations médicaux. Le RIM (Reference Information Model) introduit dans la version 3 d'HL7, est aujourd'hui l'élément de structuration conceptuelle des modèles standards proposés par les groupes de travail. Il fournit une vue statique des besoins de modélisation d'information et il permet de s'assurer de la cohérence des modélisations entre les domaines médicaux. Le processus de mise en oeuvre d'un message HL7 v3 définit des règles de dérivation des modèles d'information depuis le RIM. Ces règles demandent que toute structure d'information dérivée du RIM soit sémantiquement correcte vis à vis de celui-ci et que l'origine du modèle créé soit traçable. Le "back-bone" du RIM qui est utilisé pour définir les informations cliniques et administratives des SIH est composé de six classes :

- Classe "Act" : représente les actions ou activités de soin qui sont exécutées et qui doivent être documentées
- Classe "Participation" : définit le contexte de l'acte (qui, pour qui, où, etc.)
- Classe "Entity" : représente les éléments physiques ou des êtres participant à des activités de soins
- Classe "Role" : établit les rôles que les "Entity" jouent lors d'une activité de soin
- Classe "ActRelationship" : représente la relation d'activité de soin (causalité)
- Classe "RoleRelationship" : représente les relations entre les rôles

Le RIM est structuré suivant les relations entre ces six classes décrites dans la figure 4.1.

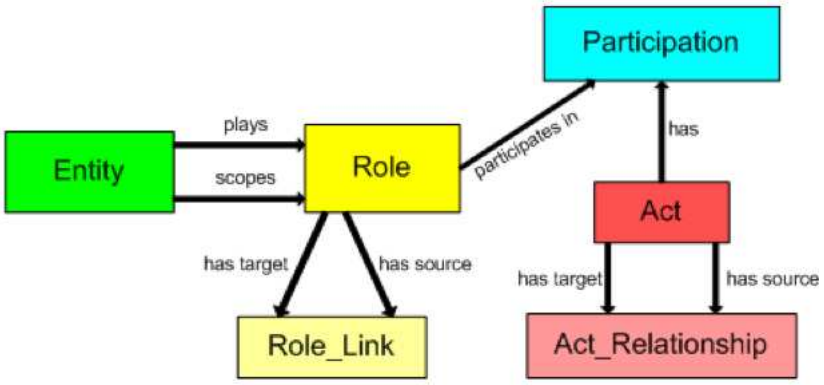


FIGURE 4.1 – Modèle conceptuel de référence d'HL7 : le RIM

On pourrait voir le RIM comme la top-ontologie d'HL7 (extrait : "In fact, if the reference information model is abstracted to a coarse level of entities and the relationships of those entities through roles to the actions that they somehow participate in then it can be conceptually applicable to any information domain or sector. One can think of a reference information model as an upper ontology that describes the static semantics of all possible real world information" [1]). En effet, définir toute information médicale enregistrable dans un SIH par ces 6 classes centrées autour d'un acte médical et du rôle d'acteurs est séduisant et permet d'offrir une représentation pivot de l'information médicale. Il reste que cette modélisation bien que louable, est critiquable [Smith 2006]. En effet, le RIM ne permet pas de faire la différence entre un élément physique et un être vivant. Prenons comme exemple le terme 'systolique'. Ce terme peut être utilisé dans un livre médical pour décrire de la connaissance à propos de la pression artérielle, et il peut être trouvé dans un rapport médical au sujet d'un acte médical effectué sur un patient. Ce sont deux informations de nature différente. La première a sa place dans une ontologie, par exemple SNOMED CT [2] . La seconde dans un modèle d'information, par exemple HL7. Le RIM ne peut faire la différence entre ces deux natures. Autre exemple, la pression artérielle est une information endurante, qui varie, et qui est continue. Alors que l'acte de mesure de cette information (l'observation) est finie dans le temps. Là encore, la distinction n'est pas aisée à faire dans le RIM. Le RIM doit donc être vu comme un élément structurant de l'information médical que HL7 v3 peut prendre en charge. Sa difficulté d'interprétation formelle additionnée du nombre de modèles UML dans le ballot HL7 en font un standard difficile à appréhender pour la communauté de l'information médicale. Il reste peu utilisé à ce jour.

4.1.2 Modèles : openEHR

openEHR est une fondation internationale qui vise à développer un standard de modélisation de l'information contenue dans les dossiers patients (EHR) [3]. openEHR est un standard ouvert permettant la modélisation et l'interopérabilité d'enregistrements éléctroniques de santé (Electronic Health Records). Tout comme HL7, openEHR définit un méta-modèle utile pour la structuration des éléments spécifiques aux domaines médicaux organisés sous forme d'archetypes. openEHR est donc organisé en deux couches distinctes : le modèle d'information et le modèle du domaine.

1. HL7 Architecture Board (ed.) : HL7 Service Aware Interoperability Framework : Canonical Version, Release 1 (Unique Ballot ID : SAIF CANON R1 I1 2011MAY). HL7, www.hl7.org/ctl.cfm ?action=ballots.home (2011)
2. Standard Nomenclature Of MEDical - Clinical Terms
3. Fondation openEHR. http://www.openehr.org/

Le CEN 13606 (standard de communication entre des dossiers patients) est l'équivalent du CDA (HL7 clinical document architecture). Ce standard définit des mappings entre openEHR et les classes Act du RIM. Le travail d'alignement n'est cependant pas encore complet.

4.1.2.1 Le modèle d'information d'openEHR

Le modèle d'information d'openEHR est un modèle abstrait représentant une vue conceptuelle de ce qu'est un dossier patient plutôt que de ce qu'est l'information clinique, le domaine. Dans openEHR, un EHR est géré comme un ensemble de *Compositions*. Ce sont les éléments de haut niveau qui sont utilisés pour enregistrer l'information à propos d'évènements ainsi que des données persistantes (par exemple : l'histoire de vaccination du patient)[Taylor 2006].

La figure 4.2 représente une vue logique de la structure du modèle d'information. Le contenu d'une composition est composé de navigation et d'enregistrement. L'enregistrement (entry) contient des informations qui doivent être enregistrées parmi les 4 types suivants : action, observation, évaluation et instruction. Un enregistrement (EHR) peut contenir plusieurs compositions qui peuvent elles-mêmes contenir plusieurs contenus qui eux-mêmes peuvent contenir plusieurs navigations et plusieurs enregistrements.

4.1.2.2 Les archétypes openEHR

Un archétype est un modèle formel d'un concept d'un domaine. À la différence d'HL7, les concepts du domaine ne sont pas simplement représentés dans une autre couche d'un modèle de classes (UML), mais comme une librairie de concepts du domaine, définie par un langage contraint. Ceci signifie que les archétypes existent pour différents enregistrements et navigations du modèle d'information openEHR. Le modèle d'information traite de la problématique de stockage de l'information du dossier patient, alors que les archétypes représentent les objets métiers que celui-ci enregistre, à différents niveaux. Voici quelques exemples d'archétypes :

- mesure du poids
- tension artérielle
- résultat microbiologique
- rapport de sortie
- prescription
- diagnostic

La "mesure du poids", la "pression artérielle" et les "résultats microbiologiques" sont des enregistrements alors que le "rapport de sortie" et la "prescription" sont

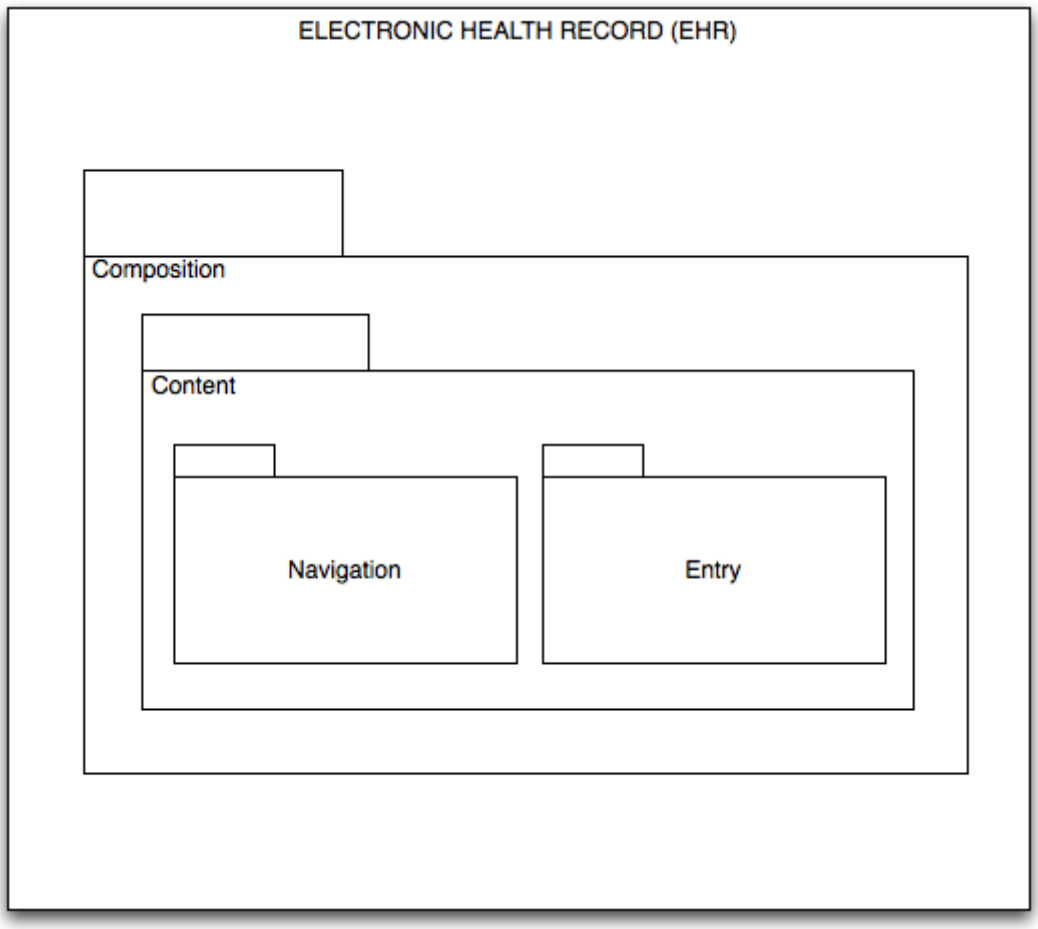


FIGURE 4.2 – Structure du modèle d'information de référence d'openEHR

des compositions. Enfin, "diagnostic" est une entrée de type évaluation.
Un archétype est généralement composé de 3 sections : entête (avec identifiant unique), la définition (contrainte sur la structure de l'archétype et cardinalités) et la section ontologie qui formalise le sens des noeuds de l'archétype ainsi que les liens vers des ressources externes comme SNOMED CT. Le langage qui permet de décrire un archétype est l'ADL ("archetype description language").
Voici un exemple d'archétype sérialisé en XML décrivant une observation de mesure de taille.

```
<OBSERVATION archetype node id=‘‘at0000’’>
  <name>
    <value>Height</value>
```

```
  </name>
  <archetype details>
    <archetype id>openEHR-EHR-OBSERVATION.height.v1</archetype id>
    <rm version>1.4</rm version>
  </archetype details>
  <data>
    <EVENT archetype node id=‘‘at0001’’>
      <name>
        <value>Any event</value>
      </name>
  <data>
      <ELEMENT archetype node id=‘‘at0002’’>
        <name>
          <value>Height</value>
          <mappings>
            <match>Length</match>
            <target>
              <terminology id>
                <value>local</value>
                <code string>at0004</code string>
              </terminology id>
            </target>
          </mappings>
        </name>
        <value>
          <units>cm</units>
          <magnitude>80</magnitude>
        </value>
    </ELEMENT>
  </data>
  </EVENT>
  </data>
</OBSERVATION>
```

4.1.2.3 Différences et complémentarités

Originellement, HL7 est un standard de message centré autour d'un acte alors que openEHR est centré autour du patient et doit plutôt être vu comme un standard

de partage de documents (EHRs). On remarquera cependant que openEHR propose maintenant son propre standard de message (EN13606) et que HL7 v3 permet des implémentations persistantes d'EHR. Dans le cadre de DebugIT les 2 standards ont été testés dans le cadre de stockage persistant et d'interopérabilité.

4.1.3 Terminologies : NEWT / ATC / ICD10

openEHR et HL7 proposent des structurations d'information (modèles d'information) standard. Cependant, le contenu de ces messages, les termes utilisés dans ces structures doivent aussi être standardisés afin d'être interopérables par une machine.
Pour ce faire, des terminologies (voir Chapitre 2) ont été créées afin d'aider au partage du contenu des messages. Par exemple,

- pour les bactéries : UniProt [4],
- pour les composés des médicaments : l'ATC (Anatomical Therapeutic Classification system) [5],
- et pour les noms des maladies : l'ICD10 (international classification for diseases v10).

Le premier niveau de l'ATC défini le groupe anatomique principal. L'ATC est donc une terminologie où les termes qui permettent de nommer les médicaments sont spécifiés, mais qui présente aussi les propriétés d'une classification suivant 5 niveaux (groupe anatomique, groupe thérapeutique principal, sous-groupe thérapeutique ou pharmacologique, sous-groupe chimique ou thérapeutique ou pharmacologique et enfin la substance chimique, le médicament). L'exemple suivant montre la classification de l'Amoxiciline dans la branche des anti-infectieux à usage systémique.

```
Antiinfectives for Systemic Use (J)
  Antibacterials for Systemic Use (J01)
    Tetracyclines (J01A)
    Amphenicols (J01B)
    Beta-Lactam Antibacterials, Penicillins (J01C)
      Penicillins with Extended Spectrum (J01CA)
        Ampicillin (J01CA01)
        Amoxicillin (J01CA04)
        Piperacillin (J01CA12)
```

4. http://www.uniprot.org/taxon
5. http://www.whocc.no/atc_ddd_index/

Il est intéressant de noter que l'Amoxiciline peut se trouver classée dans une autre branche de l'ATC. Ce qui médicalement et selon les axes de classification est normal, représente un problème lorsqu'on veut par exemple utiliser la relation de subsomption pour déduire le sous-groupe chimique auquel la substance appartient. De plus, seul le premier niveau de la hiérarchie offre une sémantique claire (organes, chimie) alors que les suivants sont un mélange de pharmacologie, de chimie et de thérapeutique.

Plus généralement, les terminologies standard présentent des structurations qui peuvent rendre leur utilisation complexe dans le cadre de partage d'information.

4.1.4 Ontologie : SNOMED CT et UMLS

SNOMED CT (Systemised Nomenclature of MEDicine - Clinical Terms) est une terminologie standard pour la santé qui contient plus de 344000 concepts actifs [6] SNOMED CT a été conjointement développé par le collège américain des pathologistes et le National Health Service au royaume-uni. SNOMED CT est aujourd'hui la propriété intellectuelle de IHTSDO (International Healthcare Terminology Standards Development Organisation) [7].

SNOMED CT est une hiérarchie supportant l'héritage multiple, constituée de concepts et de leurs attributs, chaque concept étant identifié par un SCTID (Snomed CT IDentifier). Le premier niveau de hiérarchie est constitué des 'concept groups' suivants : Clinical Findings (réstultat d'une observation clinique), Procedures (activités de soin), Body Structures (structures normales et anormales du corps), Substances (substances actives constituant d'un médicament, de nourriture, ...), Physical Objects (objets physiques naturels ou fabriqués de la main de l'homme), Events (événements qui résultent d'un accident de santé), Observable Entities (procédures ou questions qui combinées avec un résultat constituent une découverte), Qualifier Values (concepts qui ne sont nulle part ailleurs dans SNOMED CT et qui sont requis pour certains attributs, ex : open, left, right, etc.). Les attributs dans SNOMED CT sont aussi des concepts appartenant au groupe de concepts Attribute.

SNOMED CT n'est cependant pas assimilable à une réelle Ontologie puisque des redondances existent et qu'il y a plusieurs manières de représenter certains concepts. Par exemple, le concept de *fracture du fémur* est représenté sous le SCTID 71620000 alors qu'il peut aussi être représenté par les concepts de *fémur* (SCTID 182046008) et de *fracture* (SCTID 72704001).

6. United States National Library of Medicine. Unified medical language system (UMLS). http://www.nlm.nih.gov/research/umls, 2008. Last Accessed : September, 2008.

7. http://www.ihtsdo.org/

4.2 Interopérabilité Sémantique en Santé

En Janvier 2009, un rapport européen [Stroetmann 2009] a dressé un panorama de la problématique d'interopérabilité des systèmes de santé. Une définition de l'interopérabilité sémantique des systèmes de santé y est proposée : "L'interopérabilité d'un système de santé est sa capacité à :

- échanger, comprendre et agir sur des informations et de la connaissance au sujet des citoyens/patients
- au-delà des différences linguistiques et culturelles des acteurs de la santé (patients, professionnels)
- à l'intérieur et à travers le système de santé de manière collaborative"

C'est dans ce contexte que l'interopérabilité sémantique est traitée afin de faciliter le codage, la transmission et l'utilisation du sens à travers les services de soins, entre les fournisseurs, les patients, les citoyens, les autorités et la recherche. L'étendue géographique de l'interopérabilité sémantique peut être locale, régionale, nationale ou trans-frontalière. L'information échangée peut être une information au niveau du patient, mais aussi de l'information agrégée utilisée dans le cadre d'études épidémiologiques, comptables ou de recherche [8].

Dans les guides de bonnes pratiques européens [Stroetmann 2009], différents niveaux d'interopérabilité sont proposés en contraste avec l'approche "plus technique" proposée dans [Gorman 2006]

Niveau 0 :	pas d'interopérabilité
Niveau 1 :	interopérabilité technique et syntactique
Niveau 2 :	2 niveaux orthogonaux d'interopérabilité sémantique
	Niveau 2a : interopérabilité unidirectionnelle
	Niveau 2b : interopérabilité bidirectionnelle de fragments
Niveau 3 :	interopérabilité sémantique complète, permettant
	le partage du contexte et la co-opération

Prenons un exemple permettant d'illustrer les 4 niveaux d'interopérabilité proposés ici : Mr Doisneau a 56 ans, il vient de déménager pour aller vivre en Espagne. Quelques semaines après son déménagement, il tombe malade. Il arrive chez son mé-

8. La recommandation européenne, COM(2008)3282 finale, définit : "Semantic interoperability means ensuring that the precise meaning of exchanged information is understandable by any other system or application not initially developed for this purpose", où "interoperability of electronic health record systems means the ability of two or more electronic health record systems to exchange both computer interpretable data and human interpretable information and knowledge", p.14. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008H0594:EN:NOT

decin généraliste local et est transféré à l'hôpital le plus proche pour des tests plus approfondis. Suivant le niveau d'interopérabilité de l'hôpital, les actions suivantes sont entreprises :

- Niveau 0 (pas d'interopérabilité) : Mr Doisneau est soumis à beaucoup de tests pour que le médecin puisse comprendre l'origine de sa douleur aigüe.
- Niveau 1 (interopérabilité technique et syntactique) : Le médecin hospitalier de Mr Doisneau peut recevoir les documents électroniques de son pays d'origine et de son généraliste local, par exemple par mail ou en se connectant sur le site web de son hôpital Français. Malheureusement, le médecin espagnol ne parle pas Français.
- Niveau 2 (interopérabilité sémantique partielle) : Le médecin espagnol peut accéder via l'internet au dossier patient de Mr Doisneau. Malgré le fait que la langue soit différente, certaines parties du rapport médical sont codées grâce à des classifications internationales et sont donc traduisibles par le système (informations démographiques, allergies, diagnostics et certaines parties de son passé médical).
- Niveau 3 (interopérabilité sémantique totale, co-opération) : Dans cette situation idéale et après qu'une identification sécurisée ait été effectuée, le système d'information espagnol est capable d'intégrer et d'analyser les informations provenant du système d'information de l'hôpital Français et du généraliste local. Aucune barrière n'existe et les informations remontées au médecin local sont équivalentes aux informations qu'il a l'habitude de traiter pour ses patients habituels. En allant plus loin même, les données anonymisées sont automatiquement remontées aux systèmes publiques et à la recherche pour un usage secondaire des données.

Cet exemple nous permet de comprendre les enjeux de l'interopérabilité dans le domaine de la santé et son utilité pour tout citoyen. Il exprime aussi l'enjeu économique de l'interopérabilité qui permet par exemple, de diminuer le nombre de tests diagnostics faits lorsque le citoyen est en déplacement, ce qui est de plus en plus courant. Enfin, de partager des avis sur des cas médicaux complexes avec d'autres spécialistes.

Dans cette section, nous allons présenter des projets de recherche portant sur le domaine de l'intéropérabilité sémantique en santé suivant les approches définies précédemment.

4.2.1 Aggrégation de modèles, approche entrepôt de données

L'approche centralisée visant à intégrer des données dans le domaine biomédical à été mis en oeuvre dans divers projets ces dernières années, parmi lesquels ATLAS[Shah 2005], BioWarehouse[Lee 2006] et BioDWH[Töpel 2008]. Le projet le plus récent et qui tend à être de plus en plus utilisé, car ouvert, est i2b2 [9].

Le projet de recherche innovant i2b2 a démarré en 2004. Il permet d'intégrer des données médicales diverses (biologie, génétique, clinique, etc.) afin de pouvoir effectuer des interrogations multicritères sur le temps (par exemple dans le domaine génétique, tenter de prédire à partir de données cliniques et génétiques le risque de survenue de polyarthrite rhumatoïde). L'architecture fonctionnelle de cet outil, orienté services (SOA [10]), propose une organisation en cellules, chaque cellule ayant une fonction spécifique au sein de l'applicatif. La cellule CRC (correspondant à l'entrepôt de données) est dédiée à la gestion du stockage, et présente une modélisation en étoile des données. Une originalité du projet concerne la définition de la cellule « Ontology Management » qui permet le stockage et l'interrogation de ressources terminologiques. Elle supporte aujourd'hui plusieurs classifications comme par exemple LOINC [11] . Elle permet la navigation dans les données stockées dans la cellule CRC selon les ressources terminologiques, chaque terme étant relié à la donnée qu'il désigne. i2b2 ne gère pas tous les types de ressources terminologiques et notamment ne prend pas en charge le typage des relations d'ontologies formelles. La cellule CRC a pour particularité d'avoir pour dimension le « concept terminologique » associé au fait mesuré, à savoir l'observation d'un patient. Une observation sera liée au concept de prélèvement biologique, et un résultat de prélèvement y sera associé comme mesure. Dans i2b2 les ressources ontologiques et les données médicales sont gérées indépendamment, par un lien bidirectionnel qui rend l'évolutivité de l'outil complexe. L'intégration d'ontologies dans des bases de données dans le but de permettre l'évolutivité pose différentes problématiques. Il est proposé dans l'outil OntoDB [Pierra 2005] d'inclure l'ontologie et un méta-modèle de l'ontologie directement dans la base de données où sont stockées les modèles physique et conceptuel de données. Cette architecture est d'ailleurs très proche de l'architecture metadata du MOF [12] (Meta Object Facility). L'utilisation d'un méta modèle permet à l'ontologie

9. Informatics for Integrating Biology and the Bedside, https://www.i2b2.org. Siteaccédéle28/08/2011.

10. Forme d'architecture de médiation ou de modèle d'interaction applicative mettant en œuvre des services

11. Logical Observation Identifiers Names and Codes

12. 4 couches de représentation des données en UML : méta-méta-modèle, méta-modèle, modèle et données.

et aux données d'être indépendantes et génériques, puisque le modèle de l'ontologie est une instance du méta-modèle. Dans cette approche, le modèle logique est créé à partir de l'ontologie, le modèle conceptuel ne pouvant évoluer que de manière simultanée avec le modèle logique des données. Il n'y a pas encore de mise en œuvre d'OntoDB dans le domaine de la santé.

4.2.2 Alignement de modèles : fédération et médiation de données

Dans le cas de l'alignement de modèles en santé, des travaux visent à proposer un modèle normalisé permettant de fédérer les différentes sources de données. HEWAF repose sur une modélisation multidimensionnelle se basant sur les classes du RIM HL7 [Stolba 2006]. D'autres projets ont été proposés. HEMSYS[Pillai 1987], TSIMMIS[Garcia-Molina 1997], BioKleisli[Davidson 1996] et TAMBIS[Stevens 2000]. La médiation peut être aidée par des ontologies, comme dans les projets MOMIS[Beneventano 2001] avec des ontologies lexicales, ou bien PICSEL[Goasdoué 1999].

Un exemple de réseau de médiation à grande échelle pour la gestion des données en cancérologie est le projet américain caBIG. Ce projet vise (comme DebugIT en Europe pour la résistance aux antibiotiques) à mettre en oeuvre une infrastructure de partage d'informations qui pourra être utile à la recherche sur le cancer. caGrid (l'infrastructure du projet, figure4.4) est une approche d'interconnection de données basée sur les modèles et orientée services. Une architecture et des protocoles de type grille sont utilisés. caGrid utilise trois *framework* de type *middleware* : Globus Toolkit (GT), OGSA-DAI et Mobius [13]. GT est utilisé pour le management des services de la grille afin de déployer, de créer et d'invoquer les services. OGSA-DAI est utilisé pour virtualiser les sources de données et les promouvoir au rang de services de données en grille. Mobius permet la gestion de la distribution des données et des métadonnées, Mobius est par exemple employé pour gérer la gestion des schemas XML représentant la structure des types de données communs au projet. L'architecture *middleware* d'OGSA-DAI [14] permet l'intégration de données réparties ayant des modèles de données et des structures de stockages hétérogènes. Ainsi, il est possible d'interconnecter dans le grid [15] des bases de données relationnelles, XML, un système de fichiers ou des données volatiles [Antonioletti 2005]. Les sources de données sont embarquées dans des services de caGrid. Chacune implémente une in-

13. http://projectmobius.osu.edu

14. Open Grid System Architecture : http://www.ogsadai.org.uk/. Site accédé le 04/10/2010.

15. (ou grille informatique) Désigne une infrastructure virtuelle constituée de ressources informatiques partagées, distribuées, hétérogènes, délocalisées et autonomes

terface de visualisation des données locales. Le fournisseur de données doit lier les éléments des datasets locaux dans des objets partagés du projet qui sont décrits sous forme de modèles et de vocabulaires par un autre élément du projet, caDSR (cancer data standards repository) et EVS (entreprise vocabulary service).

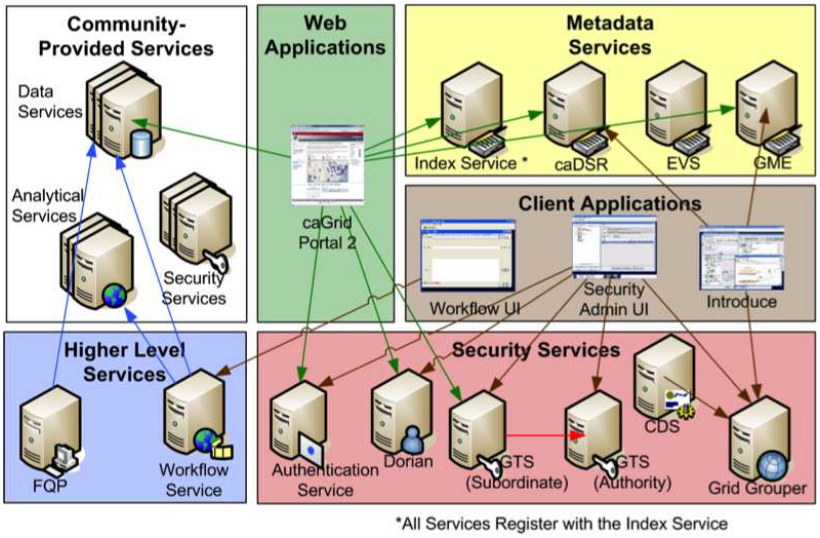


FIGURE 4.3 – Vue de l'architecture de caGrid.

La valeur ajoutée du projet caBIG est son architecture de management de l'information, à savoir les modèles de données partagés et les vocabulaires. L'interopérabilité sémantique de données biomédicales a été cependant expérimentée dans le projet caBIG au travers de leur méthodologie semCDI ([Shironoshita 2008]). Leur approche n'a pas encore pu être validée correctement à cause d'un manque de formalisation de la connaissance de leur domaine (ontologies de domaine). Cependant, des propositions d'architecture sémantique sont actuellement publiées [16] afin d'enrichir les possibilités de gestion sémantique des données de caBIG (inférence). Ceci est actuellement en cours de mise en oeuvre.

4.2.2.1 Intégration à la volée et "mashup"

L'Internet et le web de données offrent aujourd'hui des capacités d'intégration de données directement dans des applications, à la volée. Ces applications intègrent des données au moment du démarrage de l'application. Les données intégrées proviennent généralement de services web qui sont capables de distribuer des données

16. https://cabig.nci.nih.gov/2010_caBIG_Annual_Meeting_Presentations/tuesday-september-14-2010/concurrent-breakout-sessions-10-2013-14/breakout-sessions/session-13-cabigae-semantic-infrastructure-v2-overview

sur demande sur invocation du service web. Dans le domaine de la santé, quelques applications de "mashup" sont récemment apparues, comme par exemple HealthMap [17] qui vise à intégrer en temps réel une vue graphique (cartographique) de l'apparition mondiale des épidémies de maladies infectieuses ou virales. La carte intègre des données venant de différents contributeurs d'information comme "google news" ou des contributeurs locaux d'information lorsque les web services sont disponibles.



FIGURE 4.4 – Une vue des alertes d'épidémie au cours des 3 derniers mois précédent le 01/02/2011.

Il se pose évidement la question de la qualité de l'information produite et utilisée à la volée. Ce type d'application est cependant intéressante pour la mise en oeuvre de systèmes de surveillance à l'échelle d'un pays, d'un continent ou d'une région.

4.3 Analyse Comparative

Les outils étudiés peuvent être comparés selon des critères d'analyse correspondant aux différentes étapes du processus d'intégration de données à des fins d'analyse qui sont les suivants :

- l'intégration des données hétérogènes, via un processus d'intégration ETL pour un entrepôt centralisé, ou l'alignement de modèles en utilisant une approche fédérée ou de médiation ;
- la modélisation des données pour l'analyse, basée sur un modèle multidimensionnel ou relationnel, standardisé (RIM, openEHR) ou non et le stockage des données, dans des entrepôts relationnels, objets relationnels ou XML (eXist) s'il y a stockage centralisé ;

17. http://healthmap.org/fr/

– la représentation et le stockage de la dimension sémantique des données, OWL-DL, RDF, etc. et des vocabulaires ;
– l'interrogation des données stockées (par exemple via le grid si le stockage est « virtuel », via SQL (OLAP) en cas de stockage relationnel, MDX si il est multidimensionnel ou bien encore XPath/XQuery s'il est XML natif).

Nous proposons dans la figure 4.5 une analyse par critères basée d'une part sur le type d'architecture (centralisée ou décentralisée) et d'autre part sur les 4 étapes du processus d'entreposage de données 'classique'. Nous synthétisons et positionnons les systèmes étudiés en section 3 par rapport à ces critères d'analyse.

	Approche centralisée			**Approche Fédérée**	**Approche Médiée**	
Projet/Système	i2b2	OntoDB	Archimed	HEWAF	TAMBIS	caBIG
Intégration	ETL	ETL	ETL	Fédération	Mediation Ontologique	OGSA-DAI grid
Modèle d'information	Dimensionnel EAV	Relationel	Dimensionnel	Dimensionnel HL7 RIM	Ontologie globale	Modèles globaux
Prise en compte de la sémantique	Non	Oui	Non	Non	Oui	En projet
Gestion des vocabulaires	Oui	Oui	Non	Non	Oui	Oui
Langage d'interrogation	SQL + Navigateur spécifique	EXPRESS	SQL+OLAP	SQL+OLAP	Navigateur spécifique	Common Query Language

FIGURE 4.5 – Tableau comparatif des projets d'intégration de données dans le domaine de la santé (excepté pour OntoDB).

4.4 Synthèse et Discussion

Le domaine de la santé est en train de se doter de divers standards pour faciliter l'échange d'information. Ces standards sont importants et toute solution d'interopérabilité se doit de les prendre en compte. Nous mettons cependant une réserve à la standardisation, particulièrement dans le monde de l'information biomédicale. Comme nous l'avons évoqué, un standard (modèle, terminologie ou ontologie) ne représente qu'une représentation, généralement consensuelle, du monde telle qu'un groupe d'humains l'ont constitué. Cette vision ne sera que parcellaire et nécessairement diminuée par rapport à la réalité. En effet, comme nous l'avons vu dans le chapitre 2, les modèles de représentation des données ou de la connaissance induisent un appauvrissement de l'expressivité parfaite que l'on peut avoir d'un domaine. La commission Européenne, à travers ses guides de bonne pratique pour l'intéropéra-

bilité de l'information biomédicale ne décrit pas, d'ailleurs, les standards comme la solution "ultime". Le standard est bien trop souvent non interopérable avec lui-même d'une version à l'autre (ICD-9 et ICD-10). C'est pourquoi nous ne placerons pas, dans la suite de cette thèse, le débat de l'interopérabilité autour des standards. Nous aborderons ce problème scientifique en proposant une approche ancrée dans les méthodes que le web sémantique avance. À savoir de proposer un outillage permettant d'inter-relier des systèmes hétérogènes, tant du point de vue des modèles, de la sémantique, que de la qualité des données. Mais ces méthodes sont-ils suffisantes ? Sont-elles adaptables à tous les domaines ? L'interopérabilité sémantique doit au minimum résulter dans un système qui est capable de gérer l'accès aux données et leur interprétation afin de pouvoir les utiliser. L'accès aux données est quasiment résolue, et ce à grande échelle. Il reste parfois fastidieux de trouver une source de données, mais une fois trouvée, il est simple de lire l'information qu'elle présente (via des protocoles propriétaires ou web). Par contre, l'interprétation des données ainsi interrogées reste un sujet de recherche. Le web sémantique apporte des réponses grâce, notamment à l'apport de la gestion des métadonnées mais surtout de la sémantique des données. La gestion couplée données-sémantique devrait faire avancer la problématique, nous verrons si cela est suffisant dans le cadre de nos propositions dans les chapitres suivants.

Deuxième partie

Partage de l'information biomédicale dans le domaine de l'émergence de la résistance aux antibiotiques

Chapitre 5

Agents infectieux et traitements antibiotiques

"Il y a 10 fois plus de bactéries dans le corps humain que de cellules humaines."
CL. Sears [Sears 2005]

Sommaire

5.1 Introduction 92
5.2 Le contexte 92
5.2.1 L'infection bactérienne 93
5.2.1.1 Une relation comensale 93
5.2.1.2 L'infection nosocomiale 94
5.2.2 L'antibiothérapie 94
5.2.3 L'antibiogramme 95
5.2.4 L'évolution de la résistance 96
5.2.5 Les propositions pour contenir la pandémie 97
5.3 DebugIT 98
5.3.1 Introduction 98
5.3.2 Contributions attendues de DebugIT 99

La bataille contre les bactéries pathogènes est perdue d'avance. C'est une bataille contre l'évolution que l'homme seul ne peut gagner. La question reste cependant valable. Comment est-il possible de mieux réagir pour réduire ou éviter certaines pandémies ou certaines infections ? Est ce que l'outil informatique peut nous aider à mieux prescrire ? Est ce que mieux prescrire peut permettre la diminution de l'évolution de la résistance ? Qu'est ce qu'une résistance ? Nous parlerons dans ce chapitre du projet européen DebugIT et de son domaine d'application. Nous aborderons le problème de la résistance aux antibiotiques, et nous présenterons les pistes que l'Europe envisage.

5.1 Introduction

Dans la partie précédente, nous avons défini le cadre méthodologique de notre travail en rapport avec les sciences de l'information. Dans cette partie, nous allons définir dans quel cadre d'application nos recherches se situent, puis nous présenterons nos apports dans le domaine du partage d'information biomédicale, enfin, nous présenterons nos résultats et nous les évaluerons. Ce premier chapitre est dédié à la présentation du domaine informationnel dans lequel nos travaux sont expérimentés : l'étude de l'évolution de l'antibiorésistance en Europe. Puis, nous présenterons le projet Européen dans lequel nos recherches ont été effectuées.

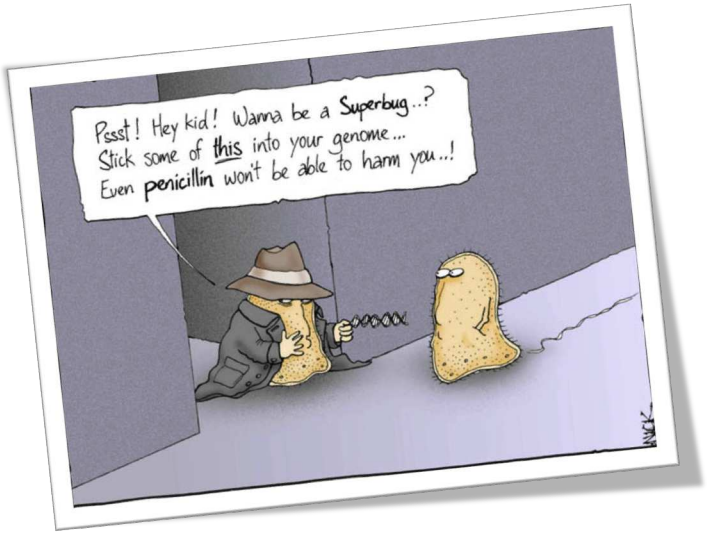


FIGURE 5.1 – Une bactérie mute et s'adapte à son milieu.

5.2 Le contexte

Après un demi siècle d'utilisation d'antibiotiques, la médecine est aujourd'hui face à un problème nouveau et grandissant : l'émergence de résistances aux antibiotiques [1]. Cette problématique est particulièrement alarmante puisque de moins en moins de nouvelles classes d'antibiotiques sont créées chaque année. Les épidémies de bactéries résistantes transcontinentales sont de plus en plus fréquentes, comme

1. Voir aussi le projet DG SANCO sur la sécurité du patient en Europe (IPSE). Qui se concentre sur les infections nosocomiales et la résistance des bactéries aux antibiotiques. Leur rapport annuel est disponible ici : http ://helics.univ-lyon1.fr/deliverables/IPSE_Annual_Report_2005_25oct06.pdf

le Staphylocoque Aureus résistant à la méthiciline (MSRA) il y a quelques années, ou plus récemment, l'Enterococci (tuberculose) résistante à la vancomycine [2]. Ou encore plus récemment, une mutation de la bactérie Escherischia Coli a causé plusieurs décès en Europe (une cinquantaine) [3], et a été provoquée par l'importation de graines d'Egypte. A un niveau plus global, le conseil Européen-Américain a par exemple publié une lettre ("EU-US Atlantic Summit 2007 Agenda") aux présidents Bush, Barroso et Merkel leur demandant de mettre en oeuvre une action basée sur 3 domaines clés, un d'eux étant de mettre en oeuvre un système global de partage de données cliniques afin d'assurer l'endiguement des maladies infectieuses.

5.2.1 L'infection bactérienne

5.2.1.1 Une relation comensale

Il y a 1 million de milliards de bactéries chez l'homme en temps normal. Nous utilisons les bactéries pour divers usages de la vie quotidienne (fromage, alcool, etc.) et nous en ingérons aussi dans nos aliments (nous ingérons aussi des antibiotiques dans nos viandes ou poissons). Les bactéries font partie de notre vie. En temps normal (non pathogénique), la relation homme-bactérie est dite de commensalisme [4]. Chacun vit normalement et apporte à l'autre ce dont il a besoin. La flore intestinale est une communauté de plus de 100 espèces bactériennes. Les bactéries sont localisées dans nos organismes de manière différente. *Lactobacillus* est la bactérie principale de la flore jusqu'à ce que l'enfant prenne une alimentation diversifiée, puis *Escherichia Coli* prédomine dans l'iléon terminal, et la flore anaérobie apparaît dans le colon. Trouver *Escherichia Coli* dans les urines, par exemple, conduit au diagnostic d'infection urinaire. Les bactéries sont donc responsables de certaines pathologies en fonction de leur localisation.

2. Cf. Tackling tuberculosis - Forgotten, but not gone, in : The Economist, March 24th 2007 : http://www.economist.com/science/displaystory.cfm?story_id=8909008

3. http://www.lepost.fr/article/2011/06/03/2514022_bacterie-e-coli-dix-cas-suspects-en-france.html

4. Le commensalisme est une variante du parasitisme ; si l'hôte fournit une partie de sa propre nourriture au commensal, il n'obtient en revanche aucune contrepartie évidente de ce dernier (la relation est à bénéfice non-réciproque). Le commensalisme est une association non-destructrice pour l'hôte (ce qui le différencie du parasitisme) ; ce dernier peut tout à fait continuer à vivre et évoluer en présence du commensal et, le plus souvent, « ignore » tout de la relation. Les survies des deux organismes sont interdépendantes. - *wikipédia*

5.2.1.2 L'infection nosocomiale

« La France détient, en Europe, le record du taux de résistance aux antibiotiques, soit 50% pour la pénicilline et 28% pour la méticilline utilisées respectivement contre le pneumocoque et le staphylocoque doré, qui constituent les principales bactéries à l'origine des infections nosocomiales. » [5]

L'infection nosocomiale est une infection d'origine bactérienne ou virale acquise au sein d'un établissement de santé. L'hôpital met en présence des individus sains et de nombreux patients présentant des pathologies variées dans un milieu clos où chaque acteur se déplace, est en contact avec du matériel (poignée de porte, l'air, etc.). L'environnement hospitalier est un véritable bassin de germes où ceux-ci évoluent et s'adaptent aussi aux agents nettoyants par exemple. La sur-utilisation, ou la mauvaise utilisation, des antibiotiques sont cependant les causes les plus avérées d'infections bactériennes nosocomiales, car cela contribue fortement à la sélection des souches hospitalières multi-résistantes et qui peuvent se transmettre d'un patient à l'autre. Ces infections ne sont pas nécessairement pathogènes pour des personnes en bonne santé. « On estime qu'en France 6 à 7% des hospitalisations sont compliquées par une infection nosocomiale plus ou moins grave, soit environ 750 000 cas sur 15 millions d'hospitalisations annuelles. » [6]

Il est à noter que dans le cadre de notre cas d'étude correspondant au domaine de l'émergence de la résistance aux antibiotiques, l'infection nosocomiale est un cas particulier d'étude, nous étudierons toutes les résistances.

5.2.2 L'antibiothérapie

L'antibiothérapie est un traitement par antibiotique qui vise à réduire ou à éradiquer une colonisation bactérienne lorsqu'elle devient infectieuse, les antibiotiques sont une réponse généralement efficace lorsque l'antibiothérapie est ciblée, c'est à dire qu'un prélèvement à visée bactériologique permet de déterminer le germe responsable de l'infection. La relation qui régit la réponse à l'antibiothérapie dans le cas d'une infection bactérienne est cependant tripartite.

La figure 5.2 présente les relations entre l'hôte (par exemple l'homme), la bactérie et l'antibiotique. Elle met en avant la complexité du traitement antibiotique puisque le résultat de celui-ci est conditionné par un troisième (et essentiel) facteur, l'homme et son environnement. L'homme a une réponse immunitaire naturelle à l'infection bactérienne avérée. L'antibiotique interagit avec l'hôte pour avoir un effet sur l'infection bactérienne. La prescription d'antibiotique doit prendre en compte

5. http://www.senat.fr/rap/r05-421/r05-4212.html
6. http://www.senat.fr/rap/r05-421/r05-4211.html#toc13

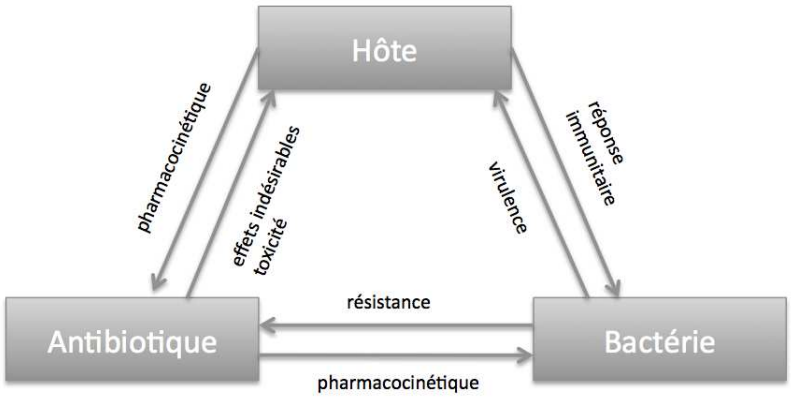


FIGURE 5.2 – Les relations entre l'hôte, la bactérie et l'antibiotique dans le cas d'une infection avérée. (Image extraite d'un document du projet DebugIT)

non seulement l'effet recherché sur l'infection des malades traités, mais aussi leurs effets sur l'écologie bactérienne et donc sur la collectivité.

5.2.3 L'antibiogramme

L'antibiogramme est une technique de laboratoire qui permet de tester la sensibilité d'une souche bactérienne à une ou plusieurs substances antibiotiques. La souche bactérienne est contenue dans un prélèvement qui est fait au patient (urines, sang, selles, salive, etc.). Ce prélèvement est ensuite mis en culture en laboratoire pour identifier les souches bactériennes présentes. Le test avec les antibiotiques est aujourd'hui effectué par un robot. Le disque sur lequel est effectué le test (mise en contact de la souche bactérienne avec plusieurs antibiotiques) révèlera le résultat de ce test en fonction du diamètre de destruction des bactéries autour de l'antibiotique. La figure 5.3 montre un résultat d'antibiogramme sur un disque. L'interprétation est différente pour chaque pays, elle consiste en la mesure du cercle blanc montrant l'efficacité d'un antibiogramme dans une culture.

L'antibiogramme permet de cibler l'antibiothérapie adéquate pour soigner un patient. En effet, certaines souches bactériennes étant résistantes à certains antibiotiques, l'antibiogramme permet d'évaluer l'état de cette résistance. La résistance est généralement interprétée suivant des seuils qui sont propres à chaque pays. Ces seuils dénotent 3 résultats possibles : Sensible (la bactérie est sensible à l'antibiotique), Intermédiaire (la bactérie est très peu sensible), Résistant (la bactérie ne réagit pas à l'antibiotique testé). Un autre indicateur permet aujourd'hui de mesurer cette ré-



FIGURE 5.3 – Un exemple de résultat d'antibiogramme : les pastilles blanches représentent un antibiotique spécifique et la matière opaque représente la culture de bactéries. Par exemple, la bactérie est sensible à CPD10 et résistante à CIP5.

sistance, c'est la concentration minimum inhibitrice (CMI). Celle-ci désigne la plus petite concentration d'antibiotique qui stoppe l'évolution de la bactérie. Son seuil d'interprétation est variable dans le temps et suivant les pays. L'antibiogramme est un outil indispensable pour mesurer l'état de l'évolution de la résistance aux antibiotiques, et bien que son utilisation soit assez commune, il reste difficile pour un médecin de modifier ses habitudes de prescription quand le patient semble guérir alors que le médicament prescrit est résistant en laboratoire.

5.2.4 L'évolution de la résistance

La résistance aux antibiotiques est devenue une pandémie globale. L'organisation mondiale de la santé (WHO) disait en 2004 : "Today we are witnessing the emergence of drug resistance along with a decline in the discovery of new antibacterials... As a result, we are facing the possibility of a future without effective antibiotics. This would fundamentally change the way modern medicine is practised." Les maladies infectieuses représentent la troisième cause de mortalité en Europe [Vicente 2006]. La résistance aux antibiotiques représente une partie majeure du problème.

En France, la prévalence de la résistance bactérienne aux antibiotiques est pré-

occupante [7]. Certaines bactéries sont maintenant totalement résistantes aux antibiotiques. Dans le même temps, le nombre d'antibiotiques mis à disposition est de plus en plus limité (peu de nouvelles molécules). Par ailleurs, cette prévalence de bactéries multirésistantes, et la gravité des infections qu'elles induisent, poussent les professionnels de santé à prescrire largement les quelques molécules encore actives ce qui va très certainement favoriser la mutation des bactéries sensibles rapidement. Récemment, des chercheurs ont montré l'émergence en Inde, au Pakistan, mais aussi au Royaume-Uni, d'un nouveau mécanisme de résistance aux antibiotiques, qui peut constituer une sérieuse menace dans la guerre contre les bactéries, en particulier face à la montée du tourisme médical. Le premier cas d'infection par une entérobactérie productrice d'une enzyme de type New Delhi métallo-beta-lactamase (NDM-1) a été identifié en 2009 par Timothy Walsh (université de Cardiff, Royaume-Uni) chez un patient suédois qui avait été hospitalisé en Inde. Le phénomène de résistances croissantes aux antibiotiques conventionnels affectait jusqu'ici surtout les bactéries de type Gram positif, ainsi appelées car elles prennent la coloration de Gram [8] lors d'un test. C'était le cas des staphylocoques dorés résistants à la méticilline et les entérocoques résistants à la vancomycine. Cependant, de plus en plus de bactéries de l'autre type, Gram négatif, sont concernées et le phénomène s'accroît plus rapidement que pour leurs homologues Gram positif. Un constat inquiétant car, "il y a moins de nouveaux antibiotiques ou d'antibiotiques en développement actifs contre les bactéries Gram négatif, et les programmes de développement de médicaments paraissent insuffisants pour fournir une couverture thérapeutique dans les dix à vingt ans".

5.2.5 Les propositions pour contenir la pandémie

En se basant sur le rapport de l'EASAC (European Academies Science Advisory Council) sur les maladies infectieuses, et en prenant en compte la recherche et les efforts de surveillance déjà en vigueur, le groupe de travail de l'EASAC publie un document en 2007 qui vise à proposer des pistes pour réduire la pandémie. Parmi celles-ci, est cité le rôle des technologies de l'information pour améliorer la surveillance et produire une réponse plus rapide à l'émergence de nouvelles résistances. On sait aujourd'hui que les bactéries ont une mémoire de résistance qui est limitée.

7. http://www.has-sante.fr/portail/jcms/c_665169/strategie-d-antibiotherapie-et-prevention-des-resist

8. La coloration de Gram doit son nom au bactériologiste danois Hans Christian Gram qui mit au point le protocole en 1884. C'est une coloration qui permet de mettre en évidence les propriétés de la paroi bactérienne, et d'utiliser ces propriétés pour les distinguer et les classifier. Son avantage est de donner une information rapide sur les bactéries présentes dans un produit ou un milieu tant sur le type que sur la forme. - *Wikipédia*

C'est à dire qu'a chaque nouvelle résistance acquise, elle en perdait une. D'un point de vue de la gestion de ces résistances, un scénario de surveillance à grande échelle pourrait aider les professionnels de santé à gérer ces résistances afin de faire "oublier" aux bactéries certaines résistances acquises.
En France, la HAS [9] préconise des recommandations aux établissements de soins afin de mieux adapter la prescription antibiotique et de mieux la cibler. La HAS recommande la mise en oeuvre de systèmes d'information connectés pour, entre autres, favoriser la mise en place de systèmes d'alerte capable de détecter l'émergence de résistances aux antibiotiques de manière coordonnée et rapide sur des bassins de population. C'est le constat que les instigateurs du projet DebugIT ont aussi fait.

5.3 DebugIT

www.debugit.eu

5.3.1 Introduction

La sécurité des patients et des citoyens n'est pas seulement liée à la réduction des erreurs médicales, mais aussi à un ensemble de conséquences complexes en relation avec le système de santé. Le développement rapide de pathogènes résistants aux antibiotiques est l'un des nouveaux risques émergents que doivent gérer les institutions ; les infections nosocomiales ne représentant qu'une partie du problème. La mise en oeuvre d'outils adaptés est une solution parmi d'autres, c'est la solution que les programmes cadre FP6 et FP7 européens ont choisi de favoriser. Diverses solutions eSanté ont déjà démontré leur impact bénéfique sur la qualité des soins. Les outils innovants actuellement en développement et mis en oeuvre dans d'autres projets européens (PSIP[Beuscart 2011], euADR[Coloma 2011], ePSOS[Traver 2011])

9. Haute Autorité de Santé : http://www.has-sante.fr

se concentrent sur les aspects d'alertes avancées qui incorporent des nouveaux outils de prédiction, détection et de surveillance. Ils appliquent des techniques de fouille de données multimédia et d'algorithmes de reconnaissance de formes, développent des nouvelles techniques d'intégration de données à partir de dossiers patients, et intègrent de la connaissance formelle du domaine et de l'aide à la décision à l'intérieur des systèmes d'information de l'hôpital. C'est dans ce cadre que le projet se situe.

5.3.2 Contributions attendues de DebugIT

DebugIT propose, construit et valide la mise en oeuvre d'un système basé sur les technologies de l'information, en s'inspirant des approches proposées et appliquées dans les domaines des sciences de la vie ou de biologie moléculaire, pour aider à renforcer la guerre contre les pathogènes infectieux. Chaque jour, de grandes quantités de données sont générées, collectées et stockées dans différents centres de soin, cependant, un volume très restreint de ces données est sollicité pour une utilisation secondaire (amélioration des soins). Le volume des données collectées va croitre avec le temps et il est aujourd'hui urgent de développer des outils qui permettront de gérer celui-ci. DebugIT va spécifiquement aborder les obstacles suivants qui freinent l'exploitation des données infectieuses :

- manque d'interopérabilité technique : l'intégration de bases de données propriétaires et hétérogènes avec des systèmes de données distribués est toujours une tâche complexe (cf. sections 2.2 et 3.3) ;
- manque d'interopérabilité sémantique : c'est un problème encore plus complexe ; différentes données de même sens doivent pouvoir être analysées conjointement (cf. sections 2.3 et 3.4) ;
- manque de transparence concernant la provenance des données : au-delà de la dimension sémantique des données, chaque analyse de données doit être capable de tracer la provenance, la qualité et la fraîcheur des données ;
- pauvre qualité des données réelles : les données réelles cliniques sont, de manière intrinsèque, de mauvaise qualité (manquantes, erreurs et bruit) (cf. section 2.4) ;
- la barrière du multimédia : connaissance et information sont liées à la capacité d'un système à regrouper différents types d'information (texte, image, son) à des fins d'analyse. Une des caractéristiques des données biomédicales est la multiplicité des formats et des types de données ;
- la confidentialité des données : la capacité d'agrégation depuis différentes sources de données ne garantit pas la vie privée des patients.

Le principal enjeu de DebugIT est de pouvoir faire communiquer des bases de données conçues pour gérer le dossier patient (pour la gestion clinique d'événements) et ce, à travers 7 fournisseurs de données européens :

- HUG : Hôpitaux Universitaires de Genève (Suisse)
- LiU : Hôpital de Linkoping (Suède)
- UKLFR : Hôpital de Freiburg (Allemagne)
- INSERM : Hôpital Européen Georges Pompidou (France)
- TEILAM : Hôpital de Lamia (Grèce)
- GAMA : Hôpital de Sofia (Bulgarie)
- IZIP : Hôpital de république Tchèque

Les données issues de ces différents centres sont très hétérogènes du point de vue de la structure et des vocabulaires qu'elles fournissent mais aussi, différentes en terme de granularité, et de qualité. C'est dans ce cadre de partage d'information biomédicale que les équipes de DebugIT ont travaillé depuis le début du projet en Janvier 2008. DebugIT vise à proposer un système de type "boucle vertueuse" (figure 5.4). Les données issues des divers fournisseurs de données européens sont extraites de manière compréhensible et sont envoyées aux "data miners" pour traitement (processus de fouille de données). Les "data miners" extraient de nouvelles connaissances qui seront stockées dans l'entrepôt de connaissances (Knowledge Repository) où les connaissances issues des données seront alignées aux données issues de la littérature. Enfin, la connaissance générée peut être utilisée dans le système d'information de l'hôpital (Clinical System) pour la mise en oeuvre d'un système de surveillance, ou bien d'un système d'aide à la prescription, par exemple. La communication entre les différents sous-systèmes de DebugIT se fait grâce à des outils issus de la communauté du web sémantique (web of reasoners).

Les partenaires du projet DebugIT sont les suivants :

- Agfa HealthCare, Belgium, co-ordinateur
- University Hospital of Geneva, Suisse
- University of Geneva, Suisse
- Linköping University, Suède
- Empirica, Allemagne
- University College of London, Grande-Bretagne
- Institut National de la Santé et de la Recherche Médicale, France
- University Medical Center Freiburg, Allemagne
- Technological Educational Institute of Lamia, Grèce
- Internetovy Pristup Ke Zdravotnim Informacim Pacienta, République Tchèque
- Gama Sofia Ltd., Bulgarie

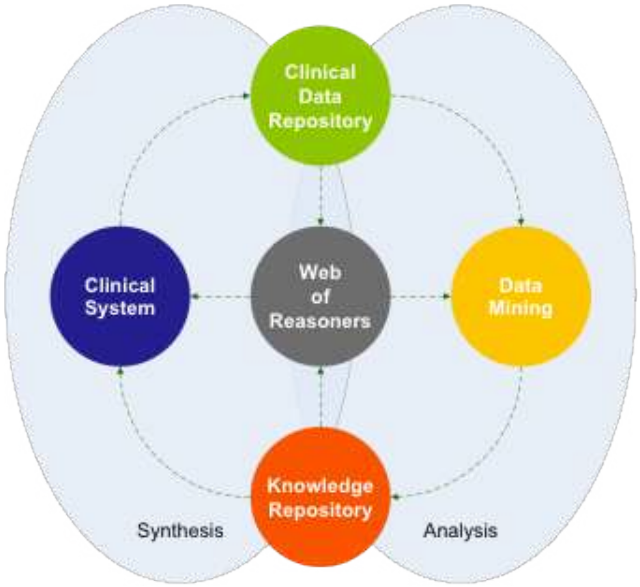


FIGURE 5.4 – La boucle DebugIT

– Averbis, Allemagne

Notre travail de thèse se situe principalement sur la partie (Clinical System - Clinical Data Repository - Web of Reasoners). C'est dans ce cadre que nous proposons des travaux sur la qualité de l'information, sur la réalisation de modèles d'information pour la mise en oeuvre d'entrepôts de données cliniques, et enfin, que nous proposons une plateforme d'interopérabilité utilisant les outils du web sémantique pour le partage d'informations biomédicales grâce à la mise en oeuvre de règles d'alignement entre des ontologies de données et une ontologie de domaine.
Le projet DebugIT devra mettre en oeuvre un socle technologique de partage d'informations biomédicales. Il utilisera les outils du web sémantique pour formaliser et partager les données à grande échelle. La plateforme issue du projet DebugIT devra permettre l'interrogation à grande échelle de données pour la surveillance de l'évolution de l'antibiorésistance en Europe et pour l'analyse de données. Diverses questions de recherches sont alors posées :

– Comment utiliser des ontologies de domaine pour interroger et exploiter des données stockées dans des bases hétérogènes et non formelles ? (robustesse de l'inférence 'métier', caractère implicite d'une base de données, médiation, montée en charge)
– Comment formaliser des bases de données biomédicales ? (modèles d'informations et ontologies, terminologies, qualité de données)
– Comment matérialiser la relation base de données - ontologie de domaine dans

un environnement multi-sites ?
- Peut-on gérer le multilinguisme, plusieurs terminologies et plusieurs représentations de l'information à la volée au moment de l'interrogation ?
- Est-ce que le modèle RDF permet l'exploitation de jeux de données volumineuses dans un cadre d'analyse ?
- Peut-on utiliser le raisonnement pour inférer la relation données - ontologie de domaine ?

Le chapitre suivant présentera nos contributions scientifiques dans le cadre de l'échange de données biomédicales ainsi que nos propositions de travaux dans le cadre du projet DebugIT.

Chapitre 6

Connaître et partager de l'information biomédicale

"Il n'y a pas une méthode unique pour étudier les choses." - Aristote.

Sommaire

6.1 Introduction 104
6.2 Connaissances liées à la qualité de l'information 106
6.2.1 La qualité de l'information et la sémiotique 107
6.2.2 Qualité de l'information source pour l'interopérabilité 109
6.2.2.1 Evaluation 109
6.2.2.2 Alignement et Surveillance 110
6.3 Des données vers la sémantique 111
6.3.1 Interopérabilité Technique 112
6.3.2 Interopérabilité Syntaxique 113
6.3.2.1 Un modèle d'analyse standardisé 114
6.3.2.2 Une approche dimensionnelle enrichie par des ressources sémantiques 118
6.3.2.3 Normalisation des termes 119
6.3.3 Interopérabilité Sémantique 120
6.3.3.1 Data Definition Ontology 122
6.3.3.2 Métadonnées et Qualité 126
6.4 La plateforme d'interopérabilité sémantique 127
6.4.1 Introduction 127
6.4.2 Fonctionalités générales d'IP 128
6.4.2.1 DebugIT Core Ontology 129
6.4.2.2 DebugIT terminologies 130
6.4.2.3 Clinical Data Repository 132
6.4.2.4 Fouille de données 133
6.4.2.5 Aide à la décision 133
6.4.3 Cas d'utilisation 133

6.4.4 Réécriture de requêtes 136
6.4.5 Le problème du monde ouvert 140
6.4.6 Une proposition partiellement satisfaisante 143
6.5 Conclusion 144

L'information biomédicale est complexe. Complexe car elle évolue. Elle évolue avec la connaissance. Elle change en fonction de l'observateur. La machine comprend mal cela. Elle a du mal à le prendre en compte. Certains diront que c'est un problème de modélisation, d'autres de puissance de calcul. Nous savons que c'est un problème d'intelligence artificielle. Car ce que l'on veut, c'est que la machine puisse interpréter le sens des choses. Qu'elle sache qu'une donnée a été mal rentrée par l'homme, qu'elle est mal codée, que le code X est le même que le code Y dans ce contexte, que l'hôpital X ne code pas comme l'hôpital Y mais ce n'est pas grave, elle comprend qu'on parle de la même chose... A l'ère des "Big Data", des données très volumineuses, on voudrait en fin de compte, que l'informatique fasse ce travail pour nous. Nous ne résoudrons pas cela dans ce chapitre. Nous poserons cependant quelques pierres pour aider à mettre en oeuvre cette utopie.

6.1 Introduction

Les données cliniques sont gérées par des systèmes informatiques cliniques qui contiennent des banques de données où se mêlent de façon hétérogène informations structurées et non structurées. Ce système ne fournit généralement pas une vision complète et unifiée des données du patient qui se trouvent disséminées à travers différents sous-systèmes. Même dans le cas de données structurées, celles-ci ne s'appuient souvent pas sur des références sémantiques standardisées et fiables, telles qu'on peut les retrouver dans les terminologies ou les systèmes de classification. Historiquement, les systèmes informatiques cliniques ont été essentiellement tournés vers les questions administratives, vers le "reporting" et vers la gestion des remboursements (CIM10 et CCAM), négligeant ainsi largement le parcours clinique du patient. Il en découle que les informations utilisables proviennent majoritairement des facturations liées aux actes médicaux, aux interventions ou aux diagnostics, et ne reflètent que rarement le processus de réflexion et de prise de décision clinique. Le problème de la diversité des systèmes d'information clinique est bien connu en Europe où l'on

trouve un système de santé fragmenté et un paysage politico-économique médical très diversifié.

Cette situation évolue cependant rapidement grâce aux programmes nationaux de eSanté (projets ANR) qui aident à la mise à jour des systèmes cliniques avec des standards internationaux et européens. L'Europe, à travers diverses initiatives et recommandations stimule la production de ces standards. Des systèmes de nouvelle génération de représentation de la connaissance contribuent à cette évolution comme SNOMED-CT [1] [Schulz 2007, Spackman 2004] et les archétypes EHR [Beale 2002]. Ils ne sont cependant pas encore matures pour être déployés et utilisés en Europe en routine. En parallèle, des avancées significatives dans le domaine de l'interopérabilité des dossiers patients ont été proposées dans les standards HL7 [2] et ISO/EN 13606. Cependant, même si à un niveau conceptuel, l'interopérabilité des systèmes d'information est proposée, il est raisonnable de penser qu'aucune convergence à grande échelle vers des dossiers patients interopérables ne sera effective avant 20 ans.

C'est dans ce contexte de système de santé hétérogènes et répartis que DebugIT doit proposer des solutions innovantes pour appréhender ces difficultés, et que nous évoluons pour trouver des méthodes pour aider à la mise en oeuvre d'une interopérabilité sémantique des données biomédicales. La connaissance dont nous disposons au démarrage du projet DebugIT, est que les données issues des dossiers patients sont d'une part pauvre en sémantique et en qualité mais surtout d'autre part, stockées avec beaucoup de connaissance implicite, que ce soit au niveau des structures d'information, des vocabulaires ou des objets. Afin de révéler le caractère implicite de l'information stockée et que nous voulons utiliser dans DebugIT, et afin de combler le manque de sémantique des données en relation avec notre domaine, nous proposons dans DebugIT un cadre formel de sémantisation des données.

La notion de partage d'information biomédicale s'inscrit, dans le cadre de notre travail, à différents niveaux où nous avons proposé des méthodologies, ou des mises en oeuvre. Nous débuterons par une proposition au niveau de la connaissance liée à la qualité de l'information pour une utilisation dans un cadre d'analyse. Nous proposerons dans ce domaine une méthodologie d'évaluation de la qualité d'une source d'information que nous expérimenterons et validerons dans le chapitre suivant qui

1. SNOMED Clinical Terms. 2007. International Health Terminology Standards Development Organization (IHTSDO) http://www.ihtsdo.org/.

2. HL-7 Standards : http://www.hl7.org/

nous permettra, entre autres, d'aider à la normalisation des termes que nous souhaitons partager. Ensuite, nous proposerons une méthodologie de conception de modèle d'information pour l'analyse de données biomédicales se basant sur HL7 ainsi que sur des techniques issues de l'ingénierie des modèles. Dans cette même partie, nous proposerons une formalisation ontologique de la base de données créée pour en aider le partage ultérieur. Puis, nous présenterons notre apport dans la spécification de la plateforme d'interopérabilité sémantique de DebugIT où la connaissance acquise sur l'information que nous devons partager dans le cadre du projet sera utilisée pour permettre le partage d'information, à un niveau sémantique. Enfin, nous présenterons notre participation aux travaux de médiation de données grâce à des règles formelles pour lier l'information d'un domaine aux données issues des EHR. Les résultats de ces diverses expérimentations seront présentés dans le chapitre suivant.

6.2 Connaissances liées à la qualité de l'information

L'utilisation de données issues des dossiers patients informatisés, ou d'autres sources de données de l'hôpital (par exemple le système informatique de gestion des résultats d'analyse de laboratoire) pose forcément le problème de la qualité de l'information. Comme nous l'avons vu dans l'état de l'art, la qualité de l'information est tout d'abord liée au contexte d'utilisation de celle-ci. Dans un contexte de soin, l'information du dossier patient est très certainement adaptée. Le médecin utilise l'informatique pour une partie de ses tâches quotidiennes alors qu'il se consacre à son activité principale : soigner. Le système d'information lui propose des services pouvant l'aider dans ses tâches quotidiennes, comme la prescription médicamenteuse informatisée, la prise de rendez-vous, la réservation de salles, la demande et le suivi d'actes médicaux auprès de confrères dans l'hôpital, mais aussi et surtout, le codage des actes et des diagnostics pour le suivi de gestion de l'activité de l'hôpital. Toutes ces tâches n'ont pas le même impact dans le travail du professionnel de santé. Pour la gestion de la prescription des médicaments, le médecin peut entrer rapidement quelques lettres que l'infirmière comprendra dans le champ 'commentaire' et se passer de chercher dans la liste déroulante le médicament précis qu'il faut donner au patient. De même, il peut tout aussi bien se passer d'utiliser l'outil informatique pour demander qu'un médicament soit prescrit au patient. Enfin, l'infirmière peut stopper la prescription pour diverses raisons comme un problème de mauvaise réaction au médicament, sans que le système d'information en soit informé. Ces cas de figure ne sont pas des cas particuliers dans le contexte du soin, car la priorité est, et restera le soin. Les systèmes d'information tendent cependant à être de mieux en

mieux adaptés au soin, et ce faisant, permettent une seconde utilisation des données du système d'information pour des analyses épidémiologiques ou pour de la recherche clinique. Mais même dans ce cas, il est et sera toujours difficile d'avoir une information parfaite en terme de qualité et de sémantique. C'est pourquoi notre première contribution vise à mettre en oeuvre un cadre méthodologique d'évaluation de la qualité de l'information biomédicale d'une source d'information dans le cadre de sa seconde utilisation, l'analyse et le partage d'information.

6.2.1 La qualité de l'information et la sémiotique

C'est dans le contexte de l'approche sémiotique définie dans le chapitre 2.4.3 que nous proposons une lecture de la qualité de l'information stockée dans les bases de données médicales dans le cadre spécifique de l'interopérabilité des données de santé. En effet, en analysant l'information issue du système d'information de l'hôpital, nous nous sommes rendus compte que celle-ci était difficile à lire et à interpréter. Pour plusieurs raisons :

- La structure de l'information d'abord, la manière dont elle est stockée et comment la relier dans ces structures de stockage.
- Le vocabulaire de l'information ensuite, comment interpréter la valeur du champ 'nom antibiotique', y a t-il des doublons ?
- Enfin, la qualité des objets de la base de données source, la date de fin de la prescription était-elle correcte si elle était supérieure à la date de début ?

Nous nous sommes alors posés la question de la nature de l'information et de sa validité. On trouve beaucoup de littérature concernant la qualité de l'information, mais celle-ci traite majoritairement de manière séparée les critères de mesure de chaque domaine (structure, vocabulaire et objet). Aucun papier ne propose de méthode intégrée permettant d'évaluer la qualité d'une source d'information classifiée suivant les 3 axes mentionnés.
L'information stockée dans le système d'information clinique peut ainsi être définie suivant trois dimensions :

- les données, ou les instances d'objets du monde réel, sont stockées physiquement dans les bases de données de santé,
- les modèles d'information représentent des concepts et des relations (parmi d'autres propriétés) qui permettent d'organiser et de structurer l'information,
- les systèmes terminologiques en santé fournissent les termes utilisés pour désigner des concepts et des relations.

Pour sa part, l'ISO distingue les terminologies (listes de termes), les thésaurus (index et synonymes), les classifications (avec des relations génériques) ou les voca-

bulaires (avec des définitions) et les ontologies (ISO TS17117). A la différence des autres systèmes terminologiques, une ontologie peut représenter les 3 sommets du triangle (concepts, termes et instances). D'une manière générale, on utilise une ontologie dans le domaine de la santé pour représenter une formalisation du domaine (concepts) et des termes d'un système d'information clinique. Nous proposons d'effectuer une classification des mesures de qualité de la littérature en fonction de ces trois dimensions définies de la manière suivante : concepts, termes et objets. Nous ne discutons pas des rapports mouvants qu'il peut y avoir entre les objets, les termes et les concepts [Bourigault 1999]. Nous sommes dans un système d'information réel pour lequel ces rapports sont fixés par la pratique.

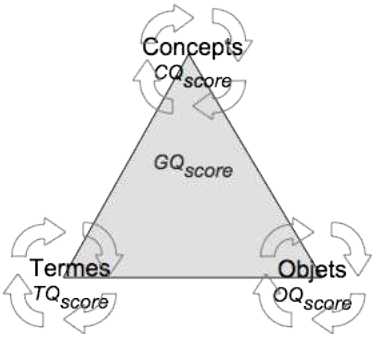


FIGURE 6.1 – Le triangle de qualité de l'information (TQI)

Les trois dimensions citées plus haut sont représentées par les sommets du triangle. Chaque sommet sera évalué et amélioré grâce à la méthode TDQM [3] ([Wang 1998]). Nous rappelons que la méthode TDQM est une méthode globale d'audit, d'évalution, d'amélioration et de surveillance de la qualité de l'information en 4 étapes. Nous proposons l'utilisation de cette méthode pour chaque sommet du triangle, tout d'abord pour en évaluer la qualité, puis pour l'améliorer et la surveiller.

Les scores de chacun des sommets peuvent être obtenus suivant différentes méthodes de la littérature ; nous en avons retenu certaines que nous avons jugées adéquates à notre domaine d'application. La moyenne des 3 scores définira un score global de qualité (GQ) qui permettra de déterminer le niveau de qualité de la source de données. La mise en œuvre de ce modèle vise à mesurer la distance entre des données et leur domaine sémantique de référence.

3. Total Data Quality Management

6.2.2 Qualité de l'information source pour l'interopérabilité

Nous avons mis en œuvre la méthodologie TDQM en quatre étapes pour évaluer le score qualité de chaque sommet. Ces quatre étapes peuvent être regroupées en deux processus complémentaires : l'évaluation (audit et qualification) de la source d'information en amont du processus de chargement des données, et l'amélioration ou l'alignement (standardisation et surveillance) lors du chargement des données dans l'entrepôt de données. Nous présentons d'abord, pour chaque sommet du triangle, les méthodes d'évaluation que nous avons utilisées, puis, les méthodes d'alignement utilisées.

6.2.2.1 Evaluation

Audit

La phase d'audit de chacun des sommets s'effectue grâce à diverses méthodes de mesure résumées dans la table 6.1. Chaque critère de qualité est généralement mesuré grâce à des méthodes algorithmiques, sauf pour la mesure de qualité du modèle d'information qui s'appuiera sur la méthode proposée par [Moody 2003b]. L'usage d'expressions régulières a par exemple été nécessaire afin de vérifier que le format des dates était uniforme. La distance terminologique est une distance statistique entre la terminologie locale et celle de référence.

Sommet	**Critère**	**Méthode**
Concepts	Domaine	Méthode subjective de mesure de qualité d'un modèle d'information
Objets	Complétude	Nombre d'enregistrements corrects sur le nombre total d'enregistrements
Objets	Précision	Adéquation du format et/ou du type des données
Objets	Unicité	Algorithme qui vérifie l'unicité des données
Objets	Cohérence	Un algorithme qui vérifie la cohérence, par exemple si la date de départ d'une prescription est inférieure à la date de fin
Termes	Cohérence	Mesure de distance aux référentiels standardisés

TABLE 6.1 – Critères et méthodes d'évaluation de la qualité de l'information

Afin de nous aider à définir notre cadre d'évaluation, nous nous tournons vers la communauté de l'informatique médicale qui a, depuis quelques années, mis en œuvre des bases de connaissances diverses comme les modèles d'information standardisés, des terminologies ou thesaurus et enfin, des ontologies [Brown 2000].

Chaque critère est mesuré en fonction d'une référence généralement consensuelle. Pour la dimension objet, les références sont des jeux de règles validées par des experts comme par exemple la date de naissance est plus récente que la date de décès. Concernant la dimension concept, nous avons utilisé les modèles d'information HL7 version 3. Pour les termes, nous avons utilisé comme référence la SNOMED CT, ICD, NEWT et WHO-ATC.

Qualification

Le processus de qualification a pour but d'établir le score de chaque dimension grâce à la phase d'audit. Nous utiliserons des degrés de qualité variant de A à D pour chaque sommet. Lorsqu'un score est un pourcentage, nous rapportons ce pourcentage à son degré correspondant (par exemple : si 73% de termes s'alignent au référentiel de termes NEWT alors la note sera B). Nous proposons l'interprétation suivante :

- A : La qualité de l'information est excellente. La source d'information est cohérente en termes de sémantique et d'organisation des données et peut être interrogée sans être adaptée.
- B : La qualité est bonne cependant il faudra améliorer la qualité d'une des dimensions du TQI.
- C : La qualité de l'information est faible. La source d'information peut être utilisée mais un effort conséquent doit être mis en œuvre pour améliorer celle-ci.
- D : La source d'information ne présente pas la qualité nécessaire pour espérer extraire de celle-ci de la connaissance et donc pour être une source potentielle de données pour un projet d'aide à la décision à partir de l'entrepôt de données.

6.2.2.2 Alignement et Surveillance

La phase de standardisation a été mise en œuvre au niveau du chargement des données depuis la source vers l'entrepôt de données de santé, TransMED, mise en oeuvre dans le cadre du projet DebugIT à l'hôpital européen Georges Pompidou (HEGP). La figure 6.2 représente la vue logique de l'architecture de mise en œuvre dans TransMED. Tout d'abord, le DPI est évalué grâce aux processus d'audit des concepts, des termes et des objets. Ensuite, lors du chargement des données et leur adaptation au modèle d'information cible (HL7), les termes sont exportés dans un référentiel de termes qui sera aligné avec les référentiels standards. Un expert

validera les termes. Lors du chargement des données, des routines permettent de contrôler continuellement le vocabulaire chargé dans l'entrepôt. Si un nouveau terme est introduit, il sera présenté à un expert si l'alignement n'est pas automatiquement fait. De cette manière, l'entrepôt de données ainsi créé présentera les caractéristiques nécessaires à l'extraction de connaissances depuis des données.

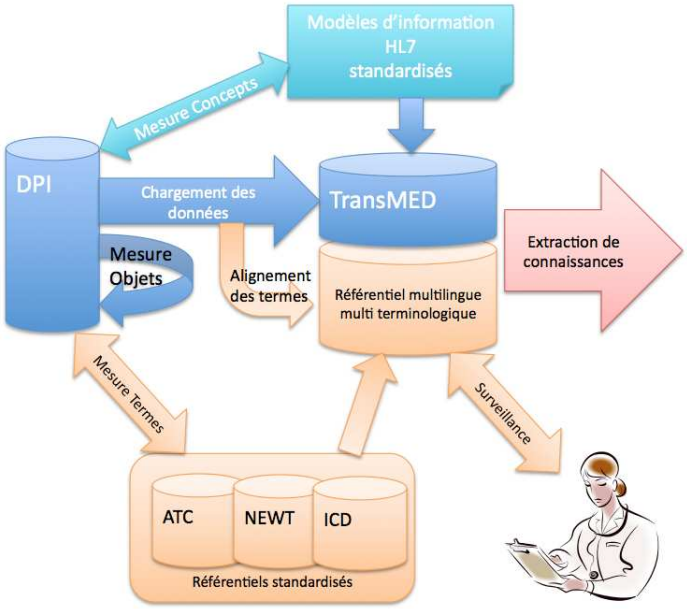


FIGURE 6.2 – Vue logique de l'architecture de qualité de TransMED.

Cette méthode a été appliquée dans le cadre de l'expérimentation à l'HEGP (Hôpital Européen Georges Pompidou) et les résultats seront présentés dans le chapitre suivant.

6.3 Des données vers la sémantique

La méthodologie d'évaluation de la qualité de l'information définie dans la section précédente nous aide à définir un cadre de normalisation et de qualité à des fins de partage de l'information biomédicale. La connaissance de la qualité de l'information n'est qu'un élément de cette connaissance nécessaire. Une autre composante est le stockage de l'information dans un modèle d'analyse interrogeable et interopérable. Dans le cadre des systèmes de santé relatifs au dossier patient, les modèles standards HL7v3 permettent de mettre en oeuvre des éléments de structure de l'information pour l'échange de messages. Des travaux concernant l'utilisation de ces modèles pour construire des bases de données sont en cours. Un des ces travaux [Ouagne 2010]

a permis la mise en oeuvre d'un modèle d'information pour une base de données basée sur HL7 dans le périmètre du projet DebugIT. Nous avons repris ces travaux et nous proposons un modèle de stockage adapté pour l'analyse d'un point de vue de la performance et de l'expressivité des requêtes que nous sommes amenés à faire sur cet entrepôt.

Mais tout d'abord, et afin d'assurer une interopérabilité sémantique au sein de la plateforme DebugIT, nous proposons une méthodologie de mise en œuvre des CDR (Clinical Data Repository) qui se décompose en 3 étapes (6.3) : technique, syntaxique et sémantique (voir Chapitre 2).

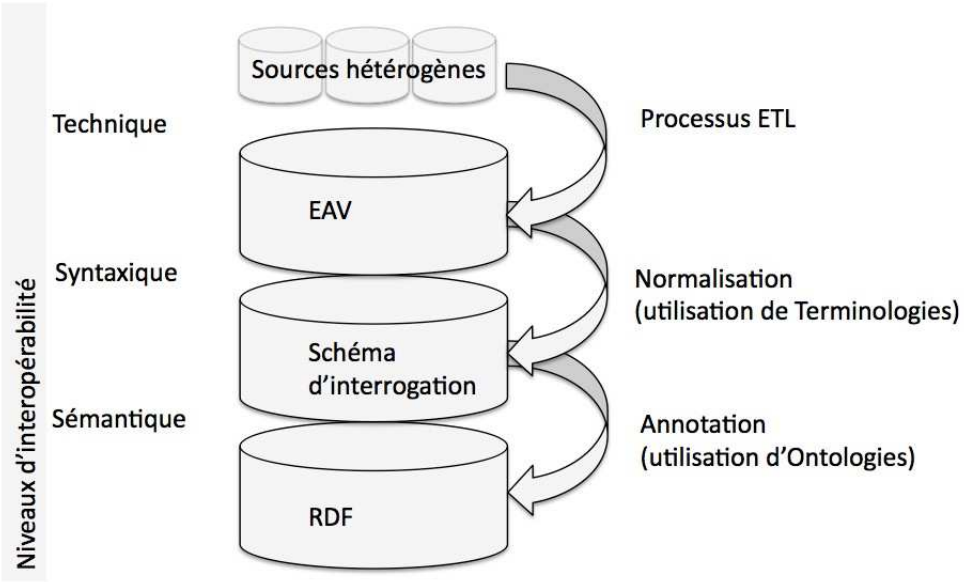


FIGURE 6.3 – Les 3 couches de l'approche d'interopérabilité pour l'intégration de données de DebugIT

6.3.1 Interopérabilité Technique

Comme décrit dans la table 6.2, les sources de données présentent différentes plateformes techniques et différents protocoles d'accès aux données. Nous proposons la mise en œuvre d'une couche technique et logique d'accès aux données à travers la mise en œuvre d'un CDR entre le système d'information hospitalier (SIH) et le réseau DebugIT. Cette couche intermédiaire de stockage de données est constituée par un espace de stockage permanent grâce à un SGBD. Le modèle de stockage proposé pour certains CDR est de la forme EAV [4]. Cette modélisation physique verticale

4. Entity-Attribute-Value : une structure de stockage de données verticale

(voir Chapitre 2.2.5) permet de diminuer le coût de traitement au niveau de la base de données lors d'opérations telles que l'ajout de nouveaux concepts ou propriétés au catalogue de données commun. En effet, la représentation des propriétés en lignes (tuples) plutôt qu'en colonnes, permet des insertions/modifications de schéma sans modifications du modèle physique de la base de données.

Source de données	**Type de stockage**	**Système d'exploitation**	**Langue**	**#épisodes de soins**
HUG	SGBD/texte libre	Windows	Français	20357
INSERM	SGBD	Windows	Français	3629551
LiU	Fichier csv	Linux	Suédois	103140
UKLFR	Texte Libre	Linux	Allemand	7949

Table 6.2 – Description des bases de données sources : vue technique

Un processus ETL est mis en œuvre entre le SIH et le CDR. Des agents d'extraction exécutent les tâches de chargement de données depuis le SIH, puis par des processus de transformation de modèle, chargent dans le CDR local. A cette étape, les sources de données DebugIT sont techniquement interopérables. Le logiciel Talend OpenStudio est utilisée pour le développement des agents d'extraction de données. Elle permet une représentation semi-automatique de la source de données (SIH) ainsi que, grâce à des modules inclus, de s'affranchir des problématiques d'accès à diverses sources de données (SGBD, XML, csv, etc.). De plus, elle permet une mise en œuvre plus aisée de la transformation de modèle SIH-CDR grâce à une interface utilisateur. Les CDR locaux sont alors mis en œuvre derrière la zone démilitarisée [5] du SI de l'hôpital.

6.3.2 Interopérabilité Syntaxique

D'une part, le modèle EAV n'est pas adapté pour interroger rapidement de grandes sources de données. Et d'autre part le contenu des sources de données (6.2) est multilingue (Français, Suédois et Allemand). De plus, des synonymes sont fréquents (ex : Staph. Aureus et Staphylocoque aureus). Chaque site ne présente pas le même niveau de normalisation par des vocabulaires ou des modèles d'information contrôlés, c'est pourquoi nous proposons à cette étape un processus de normalisation structurelle (modèle d'information standardisé), et de vocabulaires que ce soit pour

5. Représente une zone dans le SI où les données peuvent être partagées à l'extérieur de manière anonyme et sécurisée. Cette zone est généralement sur un réseau physique différent du système de production.

résoudre la problématique du multilinguisme, ou bien, la problématique de la qualité de données. La phase de normalisation des vocabulaires est par ailleurs effectuée en suivant la phase d'audit de la qualité de l'information.

6.3.2.1 Un modèle d'analyse standardisé

Une des réalisations de notre équipe de recherche a été la mise en oeuvre d'un modèle d'information physique ([Ouagne 2010]) d'une base de données pour stocker et structurer de l'information relative à la prescription antibiotique, au patient et aux résultats d'examens biologiques. Un premier modèle a été obtenu avec l'utilisation du logiciel OMDF [6] grâce auquel il est possible de récupérer des modèles UML dans le ballot d'HL7v3 en rapport avec le domaine qui nous intéresse, puis de spécialiser ces modèles conceptuels vers des modèles logiques, puis physiques. Le modèle ainsi obtenu couvre le domaine qui nous intéresse mais ne propose pas de propriétés physiques pour l'interrogation de grandes masses de données. Nous trouvons alors dans le modèle des formes non normales où différents types d'information sont stockées dans le même champ de table. Ce qui a un coût de traitement important par le moteur du SGBD [7].

Pour pallier au verrou identifié à l'issue du travail précédent, nous proposons une mise en oeuvre de modèles issus de HL7v3 instanciés vers un modèle répondant aux exigences de la modélisation dimensionnelle facilitant l'analyse. Un modèle multidimensionnel est, nous le rappelons, une variation d'un modèle relationnel qui utilise des structures dimensionnelles pour organiser l'information et exprimer les relations entre les données. Le modèle ainsi obtenu doit permettre : 1) l'exécution de requêtes complexes avec des temps d'exécution divisés par 1000 par rapport à une modélisation OLTP [8] et 2) organiser la structure de la base de données en fonction des requêtes qui seront faites, et donc de faciliter l'étape suivante qui consiste à transformer les données en RDF et les lier à une ontologie de domaine, nous y reviendrons plus tard dans ce chapitre.

Une modélisation dimensionnelle de qualité réunit les propriétés (ou recommandations) suivantes :

– Il est généralement préférable de ne pas mélanger les concepts dans une même table de fait ou de dimension.

6. https://gforge.spim.jussieu.fr/projects/omdf-hl7/
7. Système de Gestion de Base de Données
8. On Line Transaction Processing

- Une table de faits représente une action qu'il est possible de mesurer grâce à des indicateurs numériques, ou qu'il est simplement possible de compter. Une table de fait peut être généralisable (représenter un ensemble d'actions) mais il devient alors impossible de garder les propriétés d'agrégation de données du modèle multidimensionnel.
- Une dimension, ou une hiérarchie de dimension, est un ensemble de tables reliées à une ou plusieurs tables de faits, une hiérarchie de dimension traite de manière hiérarchique d'un sujet, ou d'un concept unique.
- Les relations qui lient les dimensions aux tables de faits sont contextualisantes, elles représentent la relation sémantique entre la table de fait et la dimension. Par exemple : une table de faits "Prescription Antibiotique" sera liée à la table de dimension "Antibiotique" où le nom de l'antibiotique sera renseigné. Si on veut un autre type de prescription, il faudra alors créer une autre table de faits, ou alors généraliser la table de faits "Prescription Antibiotique" à "Prescription". Il faudra alors créer une dimension "Médicament" avec une hiérarchie qui pourrait offrir une classification en fonction du type de médicament. Le problème des hiérarchies, est qu'elles ne répondent que difficilement à tous les besoins de classification.
- Il est bon pour les performances, lors de l'exécution des requêtes sur un modèle relationnel, de créer des dimensions pour les concepts qui reviennent souvent dans les requêtes. Par exemple, si on veut savoir le nombre de prélèvements urinaires dans un hôpital, il est conseillé d'avoir une dimension "Prélèvement" liée à la table de faits "Résultat Laboratoire".

Notre méthodologie de mise en oeuvre suit donc le cycle suivant : 1) Extraction des concepts depuis HL7v3 et sélection, 2) Mise en oeuvre dans l'outil OMDF d'un modèle de données logique et physique et 3) Mise en oeuvre et application des principes de développement multidimensionnels sur le modèle de données.

L'architecture de développement du modèle de données est décrite dans la figure 6.4. Nous avons particulièrement mis en oeuvre la version multidimensionnelle du modèle HL7 en suivant la méthodologie décrite précédemment.

Le matériel que nous allons utiliser est issu du ballot d'HL7. Nous avons particulièrement utilisé les modèles d'information suivants :

- A_Encounter universal (COCT_RM010000UV01)
- Result Event (PLOB_RM004000UV01)
- Composite Ordre (POOR_RM200999UV)

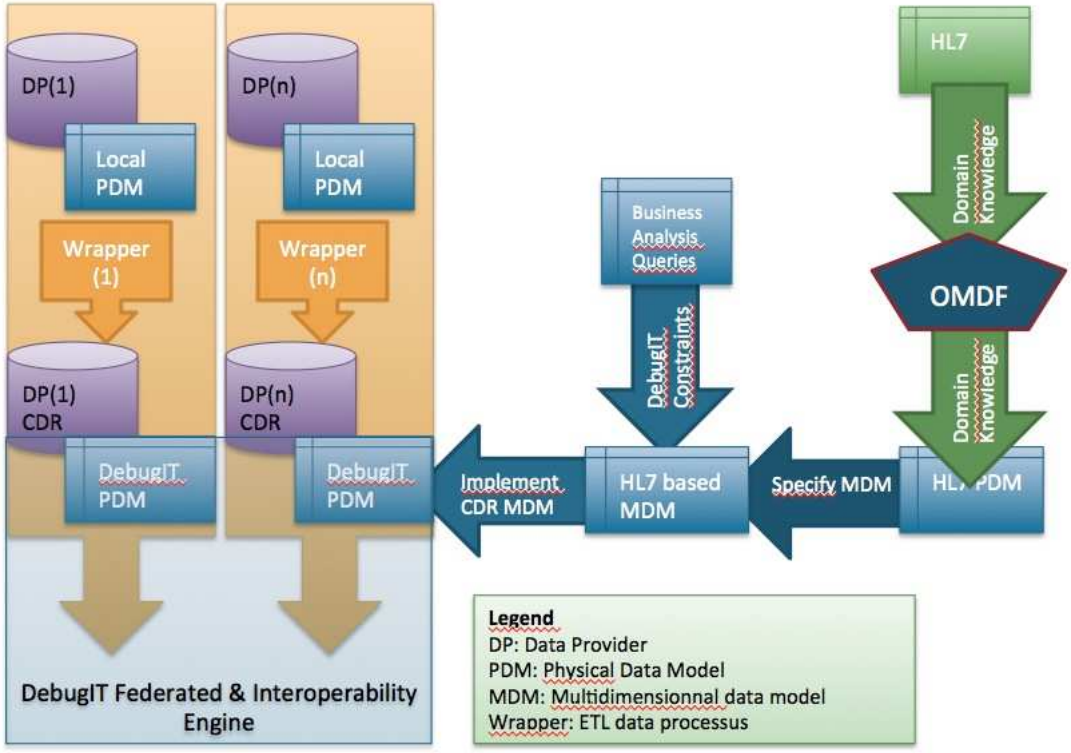


FIGURE 6.4 – Une vue du flux de travail de mise en oeuvre du modèle multidimensionnel basé sur le domaine décrit dans HL7.

- Common Observation (POOB_RM410000UV)
- Adverse Reaction (REPC_RM000022UV)
- A_BillableClinicalService Encounter (COCT_RM290004UV06)

Ces modèles d'information contiennent 61 classes et 262 propriétés. Le modèle spécialisé, puis généré dans OMDF est présenté dans la figure 6.5.

Nous mettrons en oeuvre une spécialisation multi-dimensionnelle du modèle et présenterons les résultats dans le chapitre suivant.

En parallèle de cette réalisation, nous proposons un autre travail à partir du modèle HL7 dans le cadre du projet DebugIT. En effet, les modèles HL7 contiennent de l'information appartenant au domaine lui-même, et de l'information à propos du domaine informationnel de ce même domaine (connaissance et information à propos de la même connaissance sont mélangés). Dans [Schulz 2010], il est discuté de la difficulté d'utiliser une ontologie de domaine pour de l'interopérabilité de données. La connaissance contenue dans une ontologie de domaine est relative à des objets

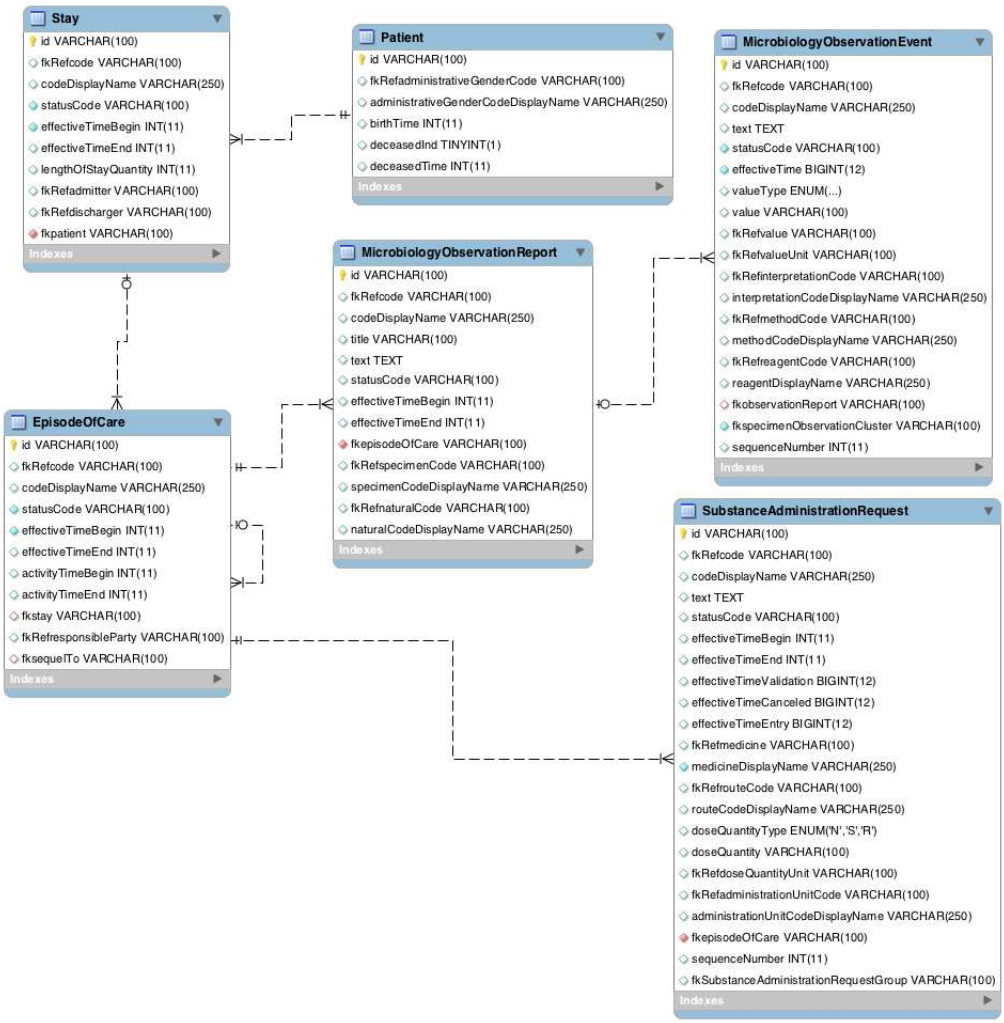


FIGURE 6.5 – Le modèle de données physique relationnel basé sur les 6 modèles d'information HL7.

matériels ou immatériels du monde. Un modèle d'information, à notre sens, ne représente que le contenant d'information à propos des objets.

Prenons l'exemple de la bactérie *E.Coli*.

Dans une ontologie de domaine pour l'étude de la résistance aux antibiotiques, E. Coli sera le plus souvent un concept. Ce concept sera associé à d'autres concepts définissant le rôle de E. Coli dans un processus de résistance. Les instances et les sous-classes de E.Coli hériteront des propriétés de E.Coli. Les instances de E.Coli dans cette ontologie représenteront les souches E.Coli testées en réalité.

Dans un modèle d'information, E.Coli ne sera pas représenté. Le concept de Bactérie sera par contre représenté. Et ce concept aura pour instance (tuple) E.Coli ainsi que

d'autres bactéries. Même lorsque le terme E. Coli est normalisé (ce qui n'est pas le cas dans HL7) et lorsque l'endroit où trouver ce terme l'est, cela n'est pas suffisant, à notre sens, pour parler d'interopérabilité sémantique (nous en reparlerons plus loin). Néanmoins, la mise en oeuvre d'un modèle d'information qui, à sa conception, prend en compte la complexité du domaine aide au rapprochement de ce modèle à son domaine. C'est dans ce cadre que nous avons participé à la mise en oeuvre d'une ontologie issue de HL7 que nous avons travaillé à intégrer à l'ontologie de domaine de DebugIT [Ouagne 2010].

6.3.2.2 Une approche dimensionnelle enrichie par des ressources sémantiques

Comme nous l'avons vu dans la première partie de cette thèse, divers modèles de représentation de l'information ont été proposés dans la littérature. Du côté des données, nous avons abordé l'évolution des représentations des modèles physiques des bases de données (relationnel, objet, dimensionnel, vectoriel, triplet). Puis nous avons présenté différents formalismes de représentation des connaissances, à travers les terminologies ou les ontologies. La particularité du modèle dimensionnel est la contextualisation des faits avec des dimensions qui peuvent représenter des hiérarchies de dimensions en exploitant la propriété de spécialisation/généralisation du lien entre les tables de dimension. La figure 6.6 exprime la sémantique du lien entre les faits et les dimensions (F est une instance de D) et entre les dimensions elles-mêmes (is_a).

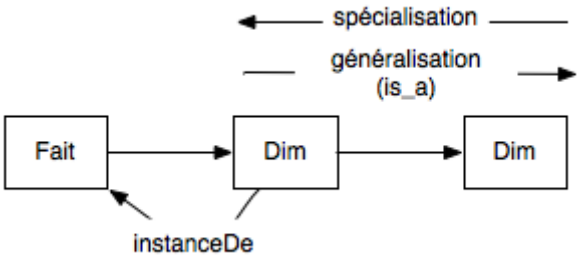


FIGURE 6.6 – Les relations conceptuelles entre les faits et les dimensions du modèle dimensionnel

Ces propriétés du modèle dimensionnel nous permettent de proposer une architecture hybride où le modèle peut être couplé par annotation à des termontologies qui implémentent des liens de hiérarchie (is_a) de même nature que ceux du modèle dimensionnel. C'est à dire, où l'agrégation des mesures contenues dans la table

des faits restent vraie. Nous proposons donc dans la figure 6.7 une architecture d'entrepôt de données permettant d'utiliser les propriétés hiérarchiques dans des formalismes de connaissances distincts des données stockées dans les formalismes dimensionnels classiques. De cette manière, la masse de données (les faits) est gérée par un moteur de SGBD relationnel performant, et la connaissance est représentée dans un formalisme expressif. Il ne restera qu'à exprimer les données dans un modèle et un formalisme compatible avec les termontologies [9] (RDF) pour être interrogées (SPARQL). La relation généralisation/spécialisation étant identique au niveau du modèle dimensionnel et de les termontologies, il est alors possible d'obtenir des résultats cohérents sémantiquement.

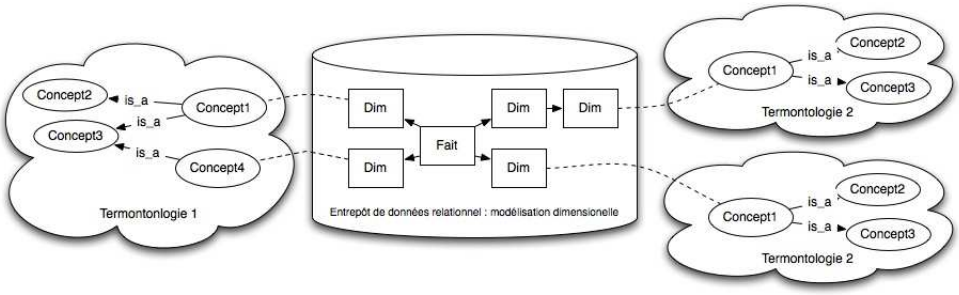


FIGURE 6.7 – Une architecture d'entrepôt de données décisionnel sémantique hybride

De plus, ce type d'architecture, permet, si l'annotation est correctement faite, de faire évoluer les classifications des termontologies sans que la base de données soit impactée. Par exemple, si un antibiotique est classé différemment et que son identifiant (URI) ne change pas, alors les résultats évolueront sans avoir à modifier à la base de données.

6.3.2.3 Normalisation des termes

Le processus de transformation des attributs codés dans un vocabulaire non contrôlé vers un vocabulaire standard et partagé est opéré par des agents de normalisation développés au niveau du CDR et au niveau de la plateforme d'interopérabilité grâce à des routines de normalisation. Certains objets comme la date sont convertis vers leur terminologie respective par les agents. Pour d'autres, comme les pathogènes ou les antibiotiques, des services tiers sont utilisés comme sources de normalisation. Dans le cas ou un concept peut prendre peu de valeurs différentes, par exemple le sexe, alors les valeurs sont liées manuellement au concept terminologique référent.

9. Termino/ontologies

Pour d'autres, comme par exemple les bactéries, un algorithme de fouille de texte simple a été mis en œuvre afin d'annoter et de normaliser les termes. Concernant les antibiotiques, l'algorithme se base sur des techniques de fouille de textes et sur un référentiel de médicaments-substance local développé par des partenaires du projet. Nous détaillerons quelques exemples de mise en oeuvre de routines de normalisation dans le chapitre suivant.
Il est cependant important de mettre en avant, d'un point de vue scientifique, le rôle de la normalisation dans l'interopérabilité sémantique. En effet, bien que les terminologies soient un élément important pour aider à interopérer des données, elles ne présentent pas toujours les propriétés nécessaires à l'annotation dans un cadre d'interopérabilité (nous avons vu l'exemple de l'ATC dans la première partie de ce mémoire). Et il reste hypothétique qu'un standard unique soit un jour trouvé et accepté pour annoter les données. Enfin, la gestion de la mise à jour et de l'interopérabilité d'une même classification reste un sujet de recherche. C'est pourquoi, même si des données ont été normalisées au niveau des CDR, nous pensons qu'il est préférable d'effectuer cet alignement de terminologies au niveau de la plateforme d'interopérabilité. En effet, au niveau de la plateforme d'interopérabilité, les données sont représentées de manière formelle en RDF. Les terminologies sont aujourd'hui représentées de la même manière. Nous pouvons à ce niveau, commencer à raisonner et à imaginer des processus d'alignement automatisés.

6.3.3 Interopérabilité Sémantique

Comme nous l'avons vu dans la première partie, plusieurs approches (bottom-up - LAV ou top-down GAV) sont proposées dans la littérature. Pour illustrer ces approches, prenons l'exemple de 2 bases de données exprimant des prescriptions antibiotiques. Afin que ces 2 bases puissent échanger de l'information, il nous faut soit une ressource pivot (ontologie ou terminologie) et aligner les 2 bases de données à cette ressource, soit des règles qui permettent d'aligner les deux bases. La figure 6.8 est une représentation graphique de cet exemple.

Bien que l'approche LAV nous paraisse plus adaptée, elle reste complexe à mettre en oeuvre. Comme nous l'avons indiqué dans la section précédente, aligner plusieurs terminologies ensemble tout en gardant une robustesse dans le temps est difficile. Nous gérons donc le lien dans DebugIT d'une part dans l'ontologie de domaine, d'autre part avec des règles de mapping dans la plateforme d'interopérabilité. Dans les deux cas, nous voyons cela comme une ressource pivot de plus haut niveau. Nous abordons donc une représentation GLAV (bottom-up et top-down) où la gestion local-global d'alignement se fera à l'aide de règles. La figure 6.9 représente notre

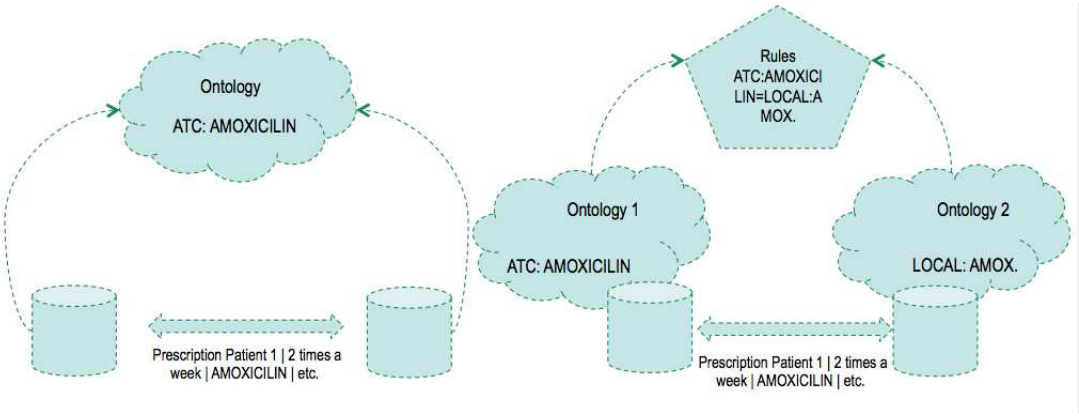


FIGURE 6.8 – A gauche : l'approche top-down où une représentation pivot est imposée pour que l'échange fonctionne, à droite : l'approche bottom-up où des règles permettent de gérer l'alignement "à la volée"

architecture de gestion d'interopérabilité sémantique.

D'un point de vue des formalismes, nous nous sommes ensuite interrogés sur la représentation la plus adaptée afin d'aider au partage de données. Après l'analyse de la littérature sur les modèles d'information en ingénierie des modèles, et sur les ontologies, comme support de stockage pour la connaissance d'un domaine, nous pensons qu'il faut élever les données au formalisme des ontologies afin de pouvoir profiter des propriétés de ces dernières. La représentation par graphes semble être la représentation adéquate dans un cadre d'interopérabilité sémantique. En effet, si nous voulons exécuter des raisonnements sur les données issues des entrepôts de données, les raisonneurs actuels prennent quasiment tous des graphes (RDF/XML, RDF/N3) en entrée. De plus, il reste plus logique d'élever vers une représentation plus riche des données, que le contraire.

Trois étapes sont définies afin de réduire la distance entre les données opérationnelles et les représentations formelles du domaine (ontologies de domaine). Premièrement, la base de données source est définie formellement (Data Description Ontology : ontologie de données) suivant trois axes : le modèle de données, vocabulaire et la qualité de données. Deuxièmement, une représentation partagée des concepts du domaine est créée ([Schober 2010]) (DebugIT Core Ontology : ontologie de domaine). Enfin, un lien entre la représentation formelle de la base source et les concepts du domaine est mis en œuvre à travers un médiateur de requêtes basé sur des règles.

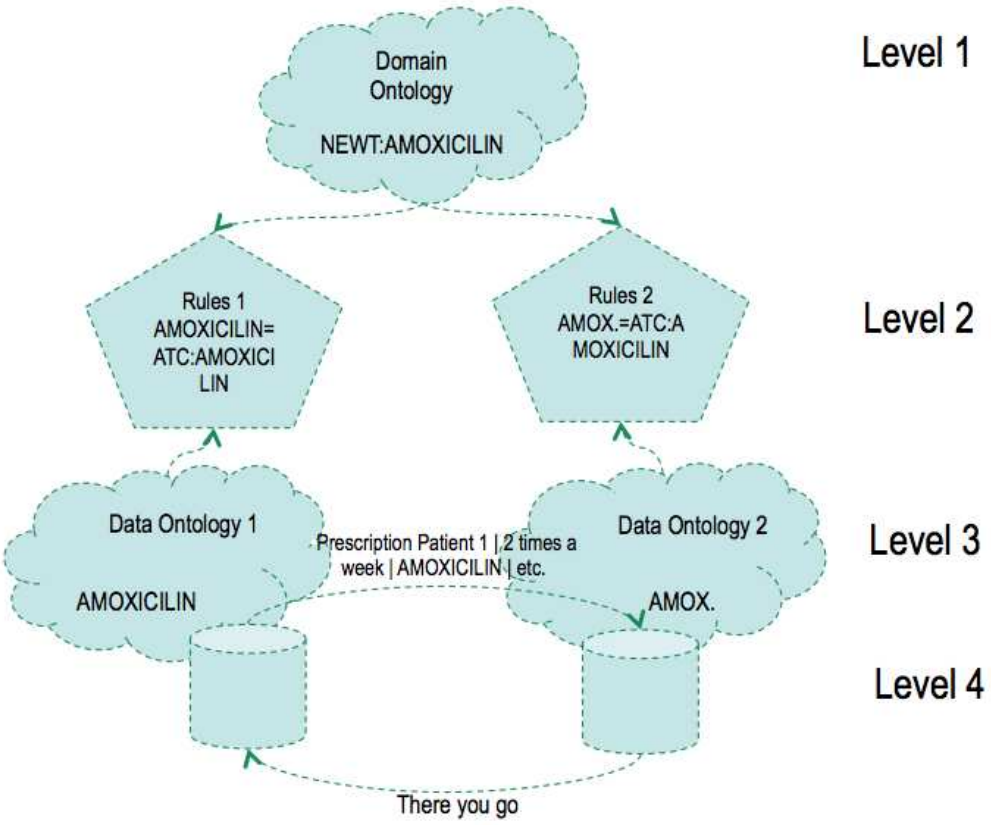


FIGURE 6.9 – Level 4 : les données au format natif, level 3 : la représentation sémantique des données (modèle d'information), level 2 : les règles d'alignement du modèle d'information formalisé avec l'ontologie de domaine, level 1 : l'ontologie de domaine, garante de la robustesse logique

6.3.3.1 Data Definition Ontology

Un nombre important d'approches pour exposer des données en RDF sur le web de données (ou LinkedData) ont été proposées dans la littérature ([Broekstra 2002] ; Openlink). Cependant, ces approches ne permettent pas de faire de l'intégration de données ([Bizer 2007]) et posent des problèmes conceptuels. En effet, il est simple de créer une vue RDF d'une base de données relationnelle grâce à des outils comme Virtuoso [10] ou D2R Server [11] où chaque table sera un concept en RDF et chaque champ de colonne sera représenté par une propriété d'un concept, et où les instances de ces propriétés seront des tuples de la base de données. Il reste cependant

10. http://virtuoso.openlinksw.com/
11. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

complexe d'aligner les concepts/propriétés/instances résultantes en RDF avec une ontologie de domaine. Ce problème d'alignement entre des données et des ontologies de domaine est étudié dans la littérature [O'Connor 2010]. Une des pistes pour résoudre ce problème est la mise en oeuvre de métadonnées sous forme d'ontologie sur les données pour aider à leur intégration avec des informations du domaine. C'est dans ce cadre que nous nous situons et que nous proposons la mise en oeuvre d'une ontologie de données sur les bases de données sources du projet.

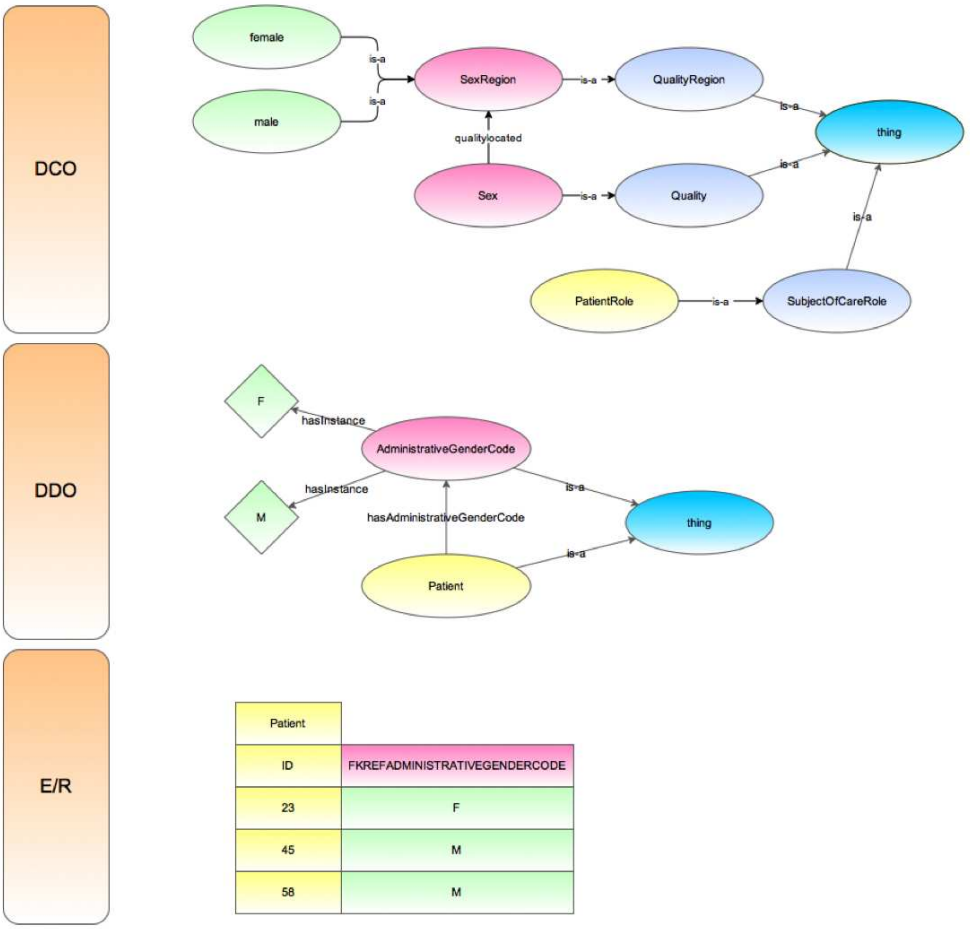


FIGURE 6.10 – 3 niveaux de représentation d'une information relative au sexe d'un patient. DCO : Ontologie de domaine (structuration héritée de BioTop), DDO : Ontologie de données, E/R : Entity-Relationshiop, modèle relationnel

La figure 6.10 représente ce problème d'alignement entre une ontologie de domaine et une base de données (la vue RDF de la base de données ne serait pas différente). Le concept de "male" est un concept dans l'ontologie de domaine, alors qu'il est une instance dans une base de données. De plus, l'organisation des concepts

de l'ontologie de domaine, dans cet exemple, ce fait en fonction d'ontologies de plus haut niveau, structurantes de celle-ci. Le sexe est soit une region soit une qualité.

Dans cet exemple, nous introduisons aussi le concept d'ontologie de données afin de fournir une représentation pivot entre la base de données et l'ontologie de domaine. Nous définissons une ontologie de données de la manière suivante :

- Une ontologie décrivant les éléments d'un modèle de données,
- où une table est un concept et où chaque concept à une URI,
- et un champ de table est un concept si celui-ci représente un concept à part entière dans notre domaine, ou bien une relation à un concept dans le cas où le champ représente une clé étrangère,
- où les relations entre les tables (clés primaires-clés étrangères) sont des relations entre des concepts et sont explicitement exprimées dans des propriétés,
- et où les vocabulaires utilisés dans les tables sont renseignés et formalisés.

Une ontologie de données est donc une représentation en OWL d'une partie des métadonnées d'une base de données, augmentée d'une information sur le vocabulaire utilisé pour certains concepts. La déclaration de classe (concept) s'effectuant de la manière suivante :

```
ddo:Concept_1
  a rdfs:Class;
    rdfs:isDefinedBy <http://ddoURI>;
    rdfs:label """table_label_EN"""@en;
    rdfs:label """table_label_FR"""@fr;
    skos:definition """concept definition english"""@en;
```

Cette définition de classe est commune à tout concept créé, et donc à toute table d'une base de données relationnelle.

Enfin, nous définissons les relations que le concept précédemment défini a :

```
ddo:Concept_1
  a rdfs:Class;
    rdfs:isDefinedBy <http://ddoURI>;
    rdfs:label """table_label_EN"""@en;
    rdfs:label """table_label_FR"""@fr;
    skos:definition """concept definition in english"""@en;
    skos:definition """concept definition in english"""@fr;
    rdfs:subClassOf [
        a owl:Restriction; owl:onProperty ddo:hasRelation_1;
```

```
        owl:someValuesFrom ddo:Concept_2], [
    a owl:Restriction; owl:onProperty ddo:hasRelation_2;
        owl:someValuesFrom xsd:int].
```

Nous remarquons ici que la deuxième (*owl Restriction*) porte sur une relation $hasRelation_2$ qui peut prendre comme valeur un ensemble de valeurs appartenant aux entiers ($xsd\,int$).

Enfin, nous proposons de définir les propriétés d'une ontologie de données associées à un concept de la manière suivante :

```
<owl:DataTypeProperty rdf:about="ddoURI/ddo.n3#hasRelation">
    <isDefinedBy rdf:resource="ddoURI/ddo.n3#"/>
    </owl:DataTypeProperty>
```

Où $ddoURI$ est l'adresse de l'ontologie de données et $hasRelation$ décrit la propriété (la relation) entre deux concepts.

Il est possible de définir un vocabulaire limité dans l'ontologie de données. Il faut d'abord définir des concepts, puis les utiliser avec la propriété $owl : oneOf$ dans le concept auquel les concepts du vocabulaire appartiennent. Par exemple, définissons un concept :

```
ddo:Concept1
    rdfs:isDefinedBy <http://ddoURI/ddo>;
    rdfs:label """label"""@en;
    skos:definition """definition"""@en;
    a ddo:Concept2.
```

Puis, nous définissons la classe auquel ce concept est attaché :

```
ddo:Concept2
    a rdfs:Class, owl:AllDifferent;
    rdfs:isDefinedBy <http://ddoURI/ddo>;
    rdfs:label """label"""@en;
    skos:definition """definition"""@en;
    owl:oneOf (ddo:Concept1 ddo:Concept3).
```

Concept3 est un concept de nature équivalente à Concept1. Le Concept2 peut donc prendre comme valeur Concept1 ou Concept3. Il est aussi possible de définir

le "range" d'un concept grâce à une (*owl Restriction*) sur un concept en indiquant que ses valeurs possibles sont de type de données "vocabulaire", comme on l'a fait avec les entiers (*xsd int*) :

```
ddo:Concept1
    a rdfs:Class;
    rdfs:isDefinedBy <http://ddoURI/ddo>;
    rdfs:label """concept term"""@en;
    skos:definition """concept definition"""@en;
    rdfs:subClassOf  [ a owl:Restriction;
     owl:onProperty ddo:hasRelation;
     owl:someValuesFrom vocabshortui:vocabdatatype].
```

Le concept "vocabshortui :vocabdatatype" désigne l'ensemble auquel le concept appartient. Celui-ci étant défini dans une ontologie qui décrit des ressources terminologiques par exemple.

Nous pensons que cette ressource est nécessaire pour la médiation de données sur le web sémantique. En effet, elle permet de décrire une source d'information telle qu'elle est vue par le concepteur de celle-ci, et elle permet de définir le type de vocabulaire que la source d'information expose. Dans le cadre de notre projet, nous utiliserons plusieurs ontologies de données pour faire de la médiation grâce à une ontologie de domaine. Lors de notre expérimentation (chapitre suivant), nous nous poserons la question de la généricité de l'approche, et nous aborderons les problèmes d'alignement entre ontologies de données et ontologie de domaine.

6.3.3.2 Métadonnées et Qualité

Comme nous l'avons vu précédemment, la qualité de l'information est un concept dynamique. Il évolue suivant l'usage que l'on veut faire des données. La mesure de qualité du champ 'prescription antibiotique' concernant la normalisation pour répondre une question relative à l'âge du patient n'aura pas le même poids que pour une question relative à la résistance aux antibiotiques. Il reste complexe d'imaginer tous les contextes d'utilisation des données, d'autant plus dans le domaine de la santé. Afin de traiter la qualité de l'information quand les données sont exposées, nous proposons d'associer aux concepts de la DDO les indicateurs qualité qui sont formellement définis dans une ontologie de la qualité. [Missier 2008] Ces indicateurs devront pouvoir renseigner l'utilisateur sur la qualité des données suivant des défi-

nitions formelles afin que celui-ci puisse juger de leur importance dans le contexte d'analyse. Nous proposons donc l'enrichissement de la DDO de la manière suivante :

```
ddo:Concept1
    a rdfs:Class;
    rdfs:isDefinedBy <http://ddoURI/ddo>;
    rdfs:label """concept term"""@en;
    skos:definition """concept definition"""@en;
    rdfs:subClassOf  [ a owl:Restriction;
     owl:onProperty ddo:hasRelation;
     owl:someValuesFrom vocabshortui:vocabdatatype];
    IQonto:hasIQMesureScore 'x';
    IQonto:hasVertex 'y'.
```

Les indicateurs ainsi exposés et définis grâce à une ontologie de la qualité peuvent permettre à l'homme et à la machine de prendre en compte la qualité d'une source d'information, dans un cadre d'interopérabilité, lors de son exploitation. Nous n'évaluerons pas cette approche dans le cadre de cette thèse.

6.4 La plateforme d'interopérabilité sémantique

6.4.1 Introduction

La plateforme d'interopérabilité de DebugIT (IP) peut être considérée comme le lien entre les différents sous-systèmes de DebugIT. Elle doit permettre l'échange d'informations sans ambiguité. Afin de lever l'ambiguité de l'information échangée, il est nécessaire d'utiliser des ressources formelles de représentation de la connaissance (Ontologies) depuis les données, jusqu'à la prise de décision. La plateforme d'interopérabilité utilise des ressources ontologiques définies dans la section précédente afin d'assurer l'intégration de données au sein de DebugIT à travers des services de normalisation, d'administration des ressources sémantiques, ou de construction de requêtes.

La plateforme d'interopérabilité, en tant que système, doit être capable de proposer des solutions d'interopérabilité sémantique dans le cadre d'échange de données au sens général. Plusieurs hypothèses sont faites concernant IP :

- Au niveau méthodologique : la plateforme utilise des connaissances pour faire communiquer des systèmes d'information. Ces connaissances sont bipartites, tout d'abord la connaissance expert du domaine modélisée dans une ou des

ontologies de domaine, et la connaissance expert informatique stockée dans une ou des ontologies de données. La plateforme doit être réutilisable dans un autre contexte (autre ontologie de domaine) et avec d'autres données (autres ontologies de données) simplement en changeant ses connaissances.

– Au niveau technologique : le web sémantique propose aujourd'hui des méthodes et outils pour la mise en oeuvre de l'interopérabilité sémantique. La plateforme devra utiliser ceux-ci et montrer leurs limites quand à l'applicabilité au domaine de la santé.

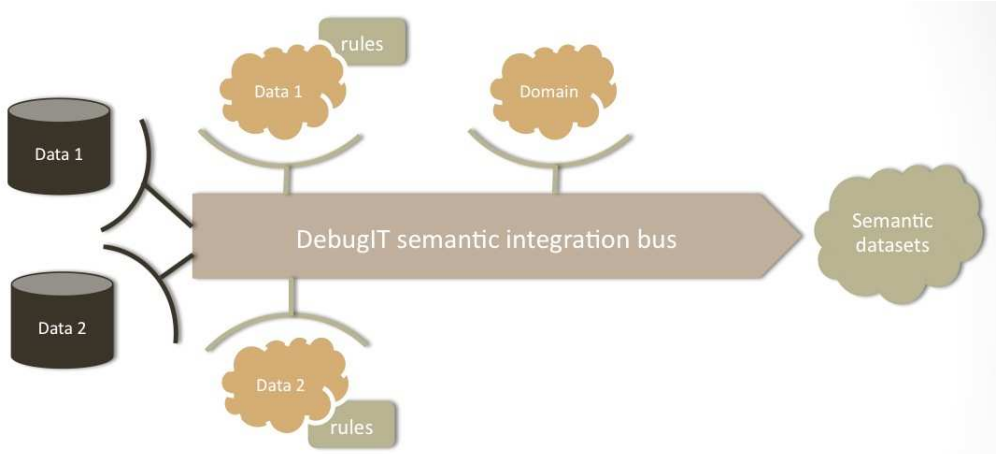


FIGURE 6.11 – La plateforme d'interopérabilité doit être réutilisable avec d'autres données, et d'autres connaissances.

La figure 6.11 présente le workflow de sémantisation des données que la plateforme d'interopérabilité doit fournir à DebugIT. Nous verrons plus loin quels services doivent être fournis pour aider à sémantiser l'information, mais nous remarquons tout d'abord que IP utilise la connaissance associée aux données, ainsi que la connaissance associée au domaine pour sémantiser des données et générer avec ceux-ci des jeux de données sémantiques. Nous insistons ici sur ce point. En effet, nous pensons que la sémantisation de données issues d'une base de données peut difficilement se faire sans représentation conceptuelle explicite de l'information d'un domaine. Il ne suffit pas de transformer une base de données en RDF pour créer de la sémantique à partir de ces données. Nous allons traiter ce problème dans cette section, ainsi que des spécifications de la plateforme d'interopérabilité, dans son ensemble.

6.4.2 Fonctionalités générales d'IP

Cette section décrit les fonctionnalités d'IP des composants logiciels qui sont inclus dans IP ainsi que d'autres services qui sont partagés entre les différents com-

posants de la plateforme. La figure 6.12 présente une vue générale des fonctionnalités d'IP en fonction des autres composants logiciels de DebugIT et des services nécessaires au fonctionnement d'IP tels qu'ils ont été présentés à la commission européenne et publiés [Choquet 2009].

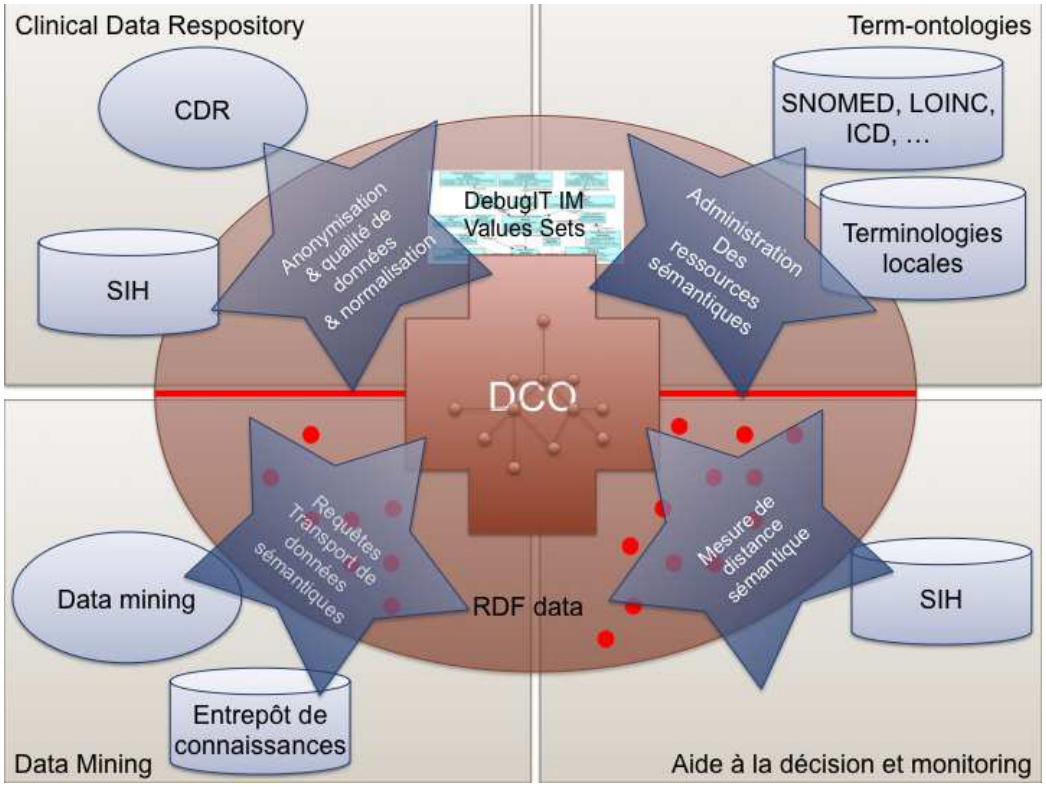


FIGURE 6.12 – La core ontology de DebugIT (DCO) est l'élément central de l'architecture, elle est utilisée par les services (étoiles) accessibles aux autres éléments de DebugIT (l'entrepôt de données cliniques, la fouille de données et l'aide à la décision)

Nous allons présenter dans cette section tout d'abord le matériel dont nous disposons pour aborder le problème de l'interopérabilité sémantique dans le cadre du projet DebugIT. Puis nous proposerons une méthode que nous expérimenterons ensuite dans le chapitre suivant.

6.4.2.1 DebugIT Core Ontology

La première ressource clé utilisée par la plateforme est l'ontologie de domaine. L'ontologie couvre l'espace conceptuel du domaine d'intérêt du projet DebugIT en particulier le domaine des maladies infectieuses au niveau clinique. L'ontologie four-

nit la référence sémantique du domaine grâce à la définition de classes et de propriétés dans un langage compréhensible par la machine. Elle représente toutes les entités nécessaires du domaine, ainsi que leurs propriétés relatives au contexte de l'étude. L'utilisation de l'ontologie dans le cadre de la plateforme d'interopérabilité encourage la mise en oeuvre dans la DCO d'identifiants sémantiques permettant d'identifier de manière unique les ressources (données) du projet. La DCO est une ontologie structurée sous BioTop et DOLCE, elle comporte à ce jour 1281 classes et 78 propriétés[Schober 2010]. L'ontologie de domaine DCO est utilisée comme l'ontologie de référence du projet. Elle n'est cependant pas la seule ontologie du projet (nous le verrons plus tard) mais elle est l'ontologie garante de la robustesse logique des inférences que l'on fera sur l'information (données) du projet. L'ontologie de domaine permet d'assurer la cohérence des concepts utilisés dans le contexte du projet, indépendamment des ressources terminologiques qui servent à annoter les données comme NEWT, l'ATC ou SNOMED CT.

La DCO est considérée comme la ressource principale du projet en terme de sémantique. Elle garantie la robustesse logique de la modélisation des concepts du domaine.
D'autres ontologies sont utilisées dans DebugIT. Ce sont les 'operational ontologies' (OO). Ces ontologies sont liées à la DCO et peuvent former une vue de la DCO pour un clinicien par exemple (Analysis and Clinical Ontologies), mais peuvent aussi permettre la formalisation d'un workflow (Workflow ontology) ou d'une requête SPARQL (SPARQL Ontology, SPARQL analysis ontology). Des ontologies permettant de gérer les unités sont aussi utilisées (Quantities and Units Extension Ontologies). Ces ontologies sont utilisées à différents moments dans le processus de médiation sémantique des données du projet, nous verrons comment dans la section suivante.

6.4.2.2 DebugIT terminologies

Lorsque cela est possible, des terminologies de référence ont été utilisées et liées à la DCO. De la même manière, ces terminologies sont utilisées pour normaliser les données au niveau des CDR (voir section suivante). Les terminologies utilisées aujourd'hui sont :

- l'ATC : classification anatomique des médicaments et substances
- la SNOMED-CT : nomenclature systématique des termes cliniques
- NEWT (UniProt) : classification universelle des protéines incluant une classification des organismes bactériens
- ICD-10 : classification internationale des maladies

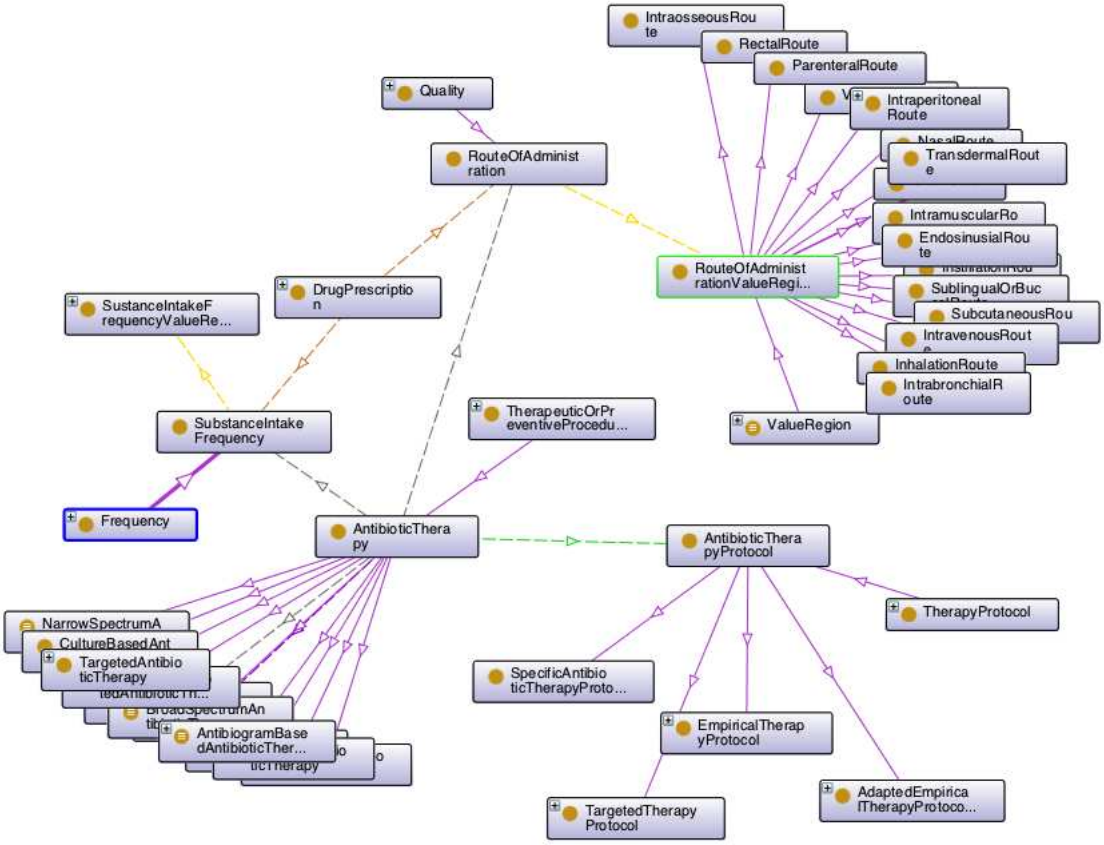


FIGURE 6.13 – Extrait de la DCO. Représentation ontologique des concepts d'antibiothérapie.

Il sera parfois nécessaire de lier des terminologies entre elles. L'IP utilisera des règles d'alignement pour gérer ce type de relation. Le vocabulaire SKOS [12] sera privilégié pour formaliser les relations inter-terminologiques. Le code suivant présente un exemple d'utilisation de SKOS pour relier des concepts SNOMED CT et ICD10.

```
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix clisko: <http://www.agfa.com/w3c/2009/clinicalSKOSSchemes#>.

# ICD10: bacterial pneumonia, unspecified; SNOMED CT: bacterial pneumonia
[skos:notation "J15.9"^^clisko:icd10DT] skos:exactMatch
[skos:notation "53084003"^^clisko:sct20080731DT].
```

12. Simple Knowledge Organization System

```
# ICD10: pneumonia due to Streptococcus pneumoniae; SNOMED CT: pneumonia due to
# Streptococcus
[skos:notation "J13"^^clisko:icd10DT] skos:broadMatch
[skos:notation "34020007"^^clisko:sct20080731DT].
```

Nous remarquons dans cet exemple que nous définissons un type de données pour les terminologies que nous utilisons. Ce type est défini dans l'ontologie "http://www.agfa.com/w3c/2009/clinicalSKOSSchemes". Ceci nous permettra d'utiliser ces ressources comme types de données dans l'ontologie de données.

6.4.2.3 Clinical Data Repository

Les entrepôts de données cliniques, sources de données principales du projet, utilisent les services de l'IP pour normaliser les données, de manière persistante ou non. IP doit aussi fournir aux entrepôts de données les composants logiciels nécessaires à l'annotation des données avec l'ontologie de domaine DCO afin de permettre la mise en oeuvre du framework logique de DebugIT. Sept fournisseurs de données sont actuellement connectés à l'IP. Ces fournisseurs de données n'ont pas d'obligation ni de se conformer à un modèle de données unique, ni à des terminologies particulières. Cependant, il est évident que le travail d'alignement est facilité lorsque les modèles où les terminologies sont synchronisées. La plateforme apporte cependant des solutions pour effectuer de l'alignement à la volée lorsque cela est nécessaire.
L'entrepôt de données pilote local que nous utilisons pour valider notre approche d'alignement sémantique est celui construit avec les données de l'HEGP. Cet entrepôt de données couvre les domaines du laboratoire, des prescriptions et des séjours des patients sur les 10 dernières années. Le modèle de données, les terminologies utilisées seront décrits dans le chapitre suivant (en tant que résultat).
Afin de modifier la syntaxe des données (relationnel vers RDF), nous utilisons l'outil D2R Server qui est un outil permettant d'exprimer en RDF le contenu d'une base de données relationnelle, et de créer par la même occasion un SPARQL endpoint qui permettra des interrogations en SPARQL sur la base. L'outil propose un langage (D2RQ) facilitant la génération de triplets RDF et l'annotation. La sérialisation du fichier de mapping est en syntaxe n3. La figure 6.14 décrit l'architecture D2R.

Nous discutons de la mise en oeuvre du fichier D2R de mapping et de l'ontologie de données qui sert à annoter ce fichier dans le chapitre suivant lors de l'expérimentation.

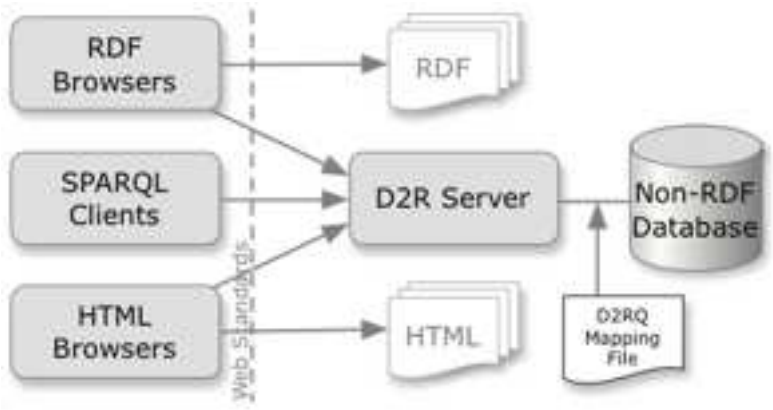


FIGURE 6.14 – L'architecture D2R : L'outil permet l'accès aux données relationnelles en format RDF en HTML, RDF et SPARQL.

6.4.2.4 Fouille de données

Les "dataminers", ou les processus de "datamining" utilisent l'IP pour créer des requêtes sur les différents CDR du projet. Un fois la requête effectuée, l'IP transforme les résultats pour les unifier sous une même représentation sémantique, la DCO. Le graphe RDF ainsi obtenu est retourné au processus de "datamining". L'IP est aussi responsable du transport des données de manière sécurisée sur l'Internet.

6.4.2.5 Aide à la décision

Des processus d'aide à la décision, incluant la génération de tableaux de bord rafraichis en temps réel sont mis en oeuvre grâce aux données générées par les CDR et les processus de "datamining". Les données sont transportées par l'IP en RDF.

6.4.3 Cas d'utilisation

Dans le cadre de cette thèse, nous nous limiterons à la présentation des cas d'utilisation concernant la médiation et l'agrégation de résultats issus des CDR. Les autres cas d'utilisations sont disponibles publiquement sur le site du projet à l'adresse `http://www.debugit.eu`

Des cas d'utilisation ont été utilisés pour définir le périmètre fonctionnel de la plateforme. Une vue générale des acteurs est présentée dans la figure 6.15.

En plus des services et ressources identifiées précédemment, nous introduisons ici les acteurs administratifs de la plateforme : Le "Core Ontology Manager" (COM) gère la connaissance associée au domaine et aide à la formalisation de la connaissance

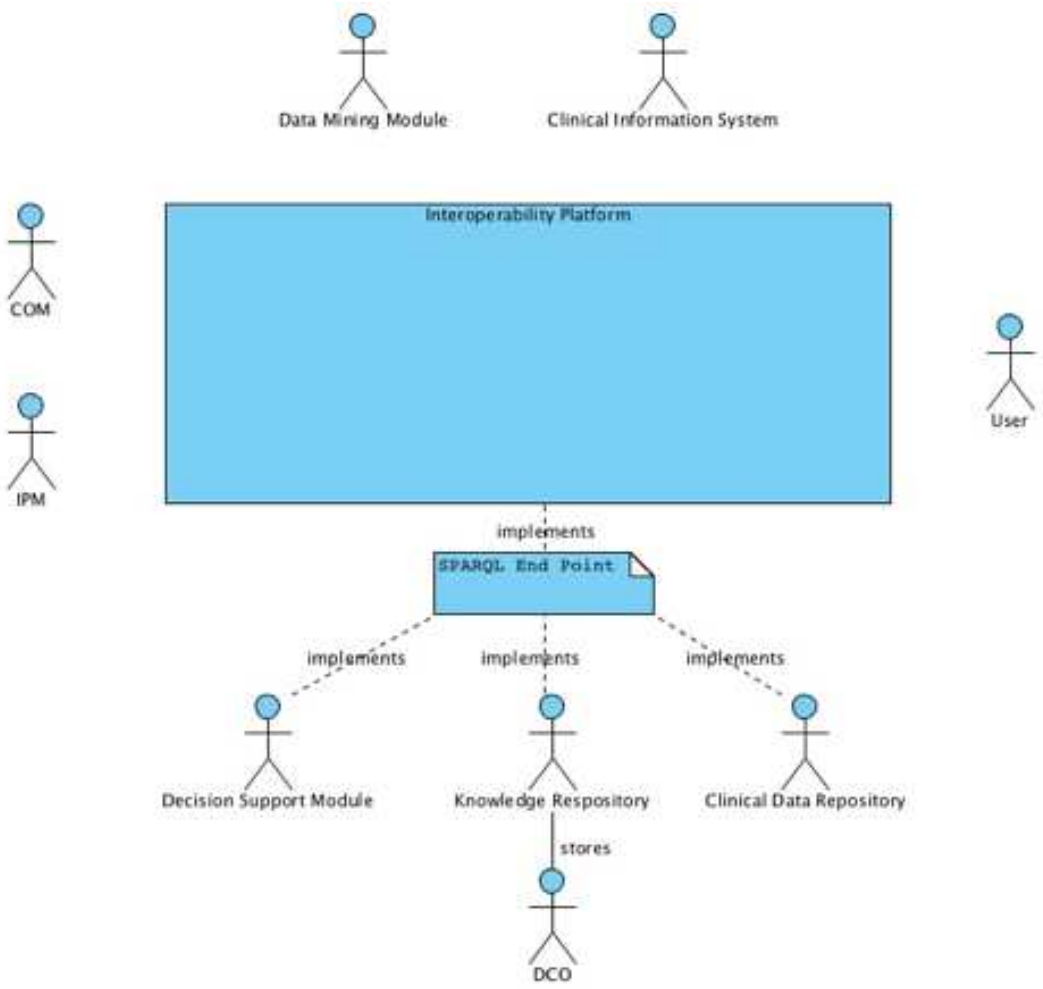


FIGURE 6.15 – Cas d'utilisation global. Représentation des acteurs de la plateforme d'interopérabilité.

associée aux données fournies par le "data manager". "L'Interoperability Platform Manager" (IPM) aide à la gestion de la plateforme, par exemple, pour l'inscription de nouveau centres de données au réseau sémantique de la plateforme.

Suivant l'approche SOA [13], chaque service de la plateforme met en oeuvre des interfaces avec d'autres services (figure 6.16).

Les acteurs actifs (services) accèdent aux services de la plateforme via des ser-

13. Service Oriented Architecture : Architecture orientée services où chaque élément logiciel peut être appelé par un autre élément logiciel, chaque élément logiciel présente une liste de services utilisables.

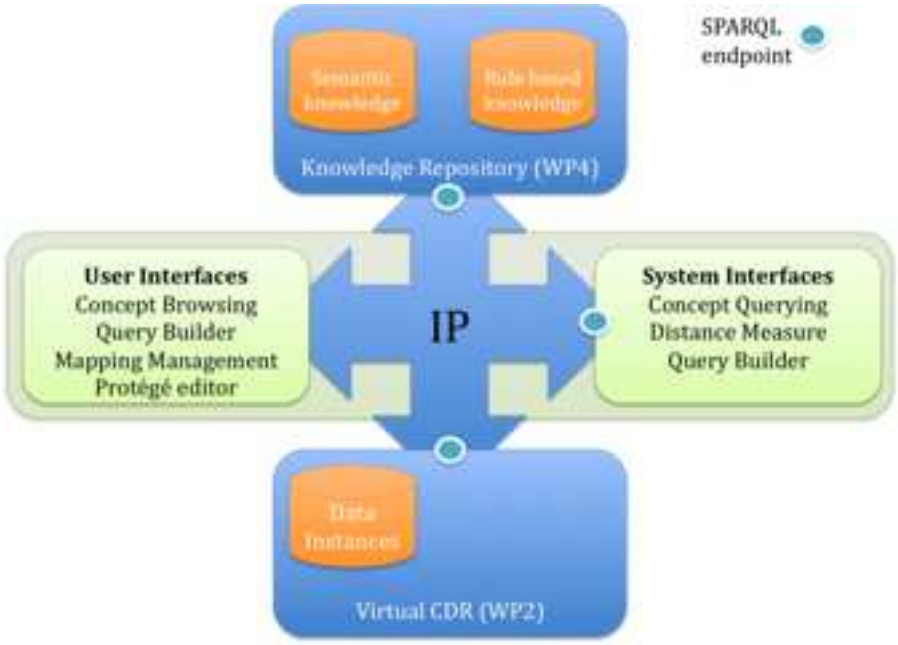


FIGURE 6.16 – Vue des interfaces d'IP et des ressources que la plateforme utilise pour opérer.

vices d'accès et de transport de l'information : SPARQL [14] alors que les acteurs passifs (humains) et administratifs accèdent à la plateforme de manière interactive, via des interfaces homme-machine le plus souvent. L'utilisation de RDF comme modèle et de SPARQL comme langage de requêtes et de transport doit se faire dans le respect des standards de transport et de modélisation du domaine de la santé. Chaque site publie des données dans des standards différents, et l'alignement de ces standards doit permettre à la plateforme de médier des informations. Les problématiques relatives à l'alignement des différents standards ne sont pas partie intégrante de la plateforme d'interopérabilité. Cependant, et nous le verrons, la plateforme doit permettre la mise en oeuvre aisée de ce type d'alignement à défaut d'avoir des alignements existants.

La plateforme se décompose en deux jeux de services pour l'interopérabilité (figure 7.4). Les services relatifs à la construction et à la gestion des requêtes, et les services liés à l'exécution et au traitement des requêtes. Dans le sous-système de gestion des requêtes sont mises en oeuvre des interfaces utilisateur permettant la construction, le stockage et l'appel de requêtes dans la base de connaissances (Knowledge Repository) qui contient les requêtes déjà formalisées. Nous capitalisons ainsi sur le travail

14. Standard Protocol and RDF Query Language - http://www.w3.org/TR/rdf-sparql-query/

de formalisation des requêtes. Le sous-système de traitement des requêtes met en oeuvre des services permettant :

- la médiation de requêtes pour l'envoi de requêtes sur différents CDR,
- la réécriture de requêtes afin d'adapter les requêtes en fonction des différents CDR (en fonction de leur ontologie de données),
- l'agrégation de données pour transformer les jeux de données provenant des CDR dans un langage interopérable (DCO),
- la sécurisation de l'accès et du transport des données par des services d'authentification et de cryptage,
- et des services d'anonymisation des données.

6.4.4 Réécriture de requêtes

Comme nous l'avons vu dans le Chapitre 3, nous nous situons ici dans une approche GLAV (global as local as view) de médiation de données où une ontologie de domaine est la ressource médiante. Nous sommes donc dans un contexte de réécriture de requête. Afin de réécrire ces requêtes, nous disposons de 2 ressources : une ontologie de domaine, décrivant les éléments du domaine et plusieurs ontologies de données, différentes d'un site à l'autre. La plateforme d'interopérabilité doit mettre en oeuvre un système de traduction devant permettre l'intégration des données issues de ces sources. La figure 6.18 présente le rôle de la plateforme d'interopérabilité vis à vis des sources de données. Nous remarquons ici que cette architecture est multi-couches, et que la réécriture de requête est faite après la mise à disposition des données où la qualité et la normalisation sont connues dans les CDRs. Puis, les données sont transformées en RDF et peuvent être interrogées par la plateforme. Afin d'assurer la réécriture de la requête principale vers les différents CDR, le projet propose la mise en oeuvre d'un moteur de réécriture à base de règles. Le processus de réécriture de requête est alors le suivant :

- L'utilisateur (le data miner) exprime sa requête dans des termes/concepts de chaque ontologie de données,
- si la requête met en oeuvre des éléments où des règles DCO->DDO n'existent pas, ils doivent être ajoutés,
- sinon, les requêtes sont exécutées sur les n CDR et les résultats sont récupérés en concepts DCO grâce à la clause CONSTRUCT du langage SPARQL
- les données sont agrégées et présentées.

Le mécanisme de CONSTRUCT de SPARQL est le mécanisme qui doit ici nous permettre d'opérer la transformation physique des concepts. En effet, celui-ci permet de construire un nouveau graphe en se basant sur un autre graphe. Par exemple :

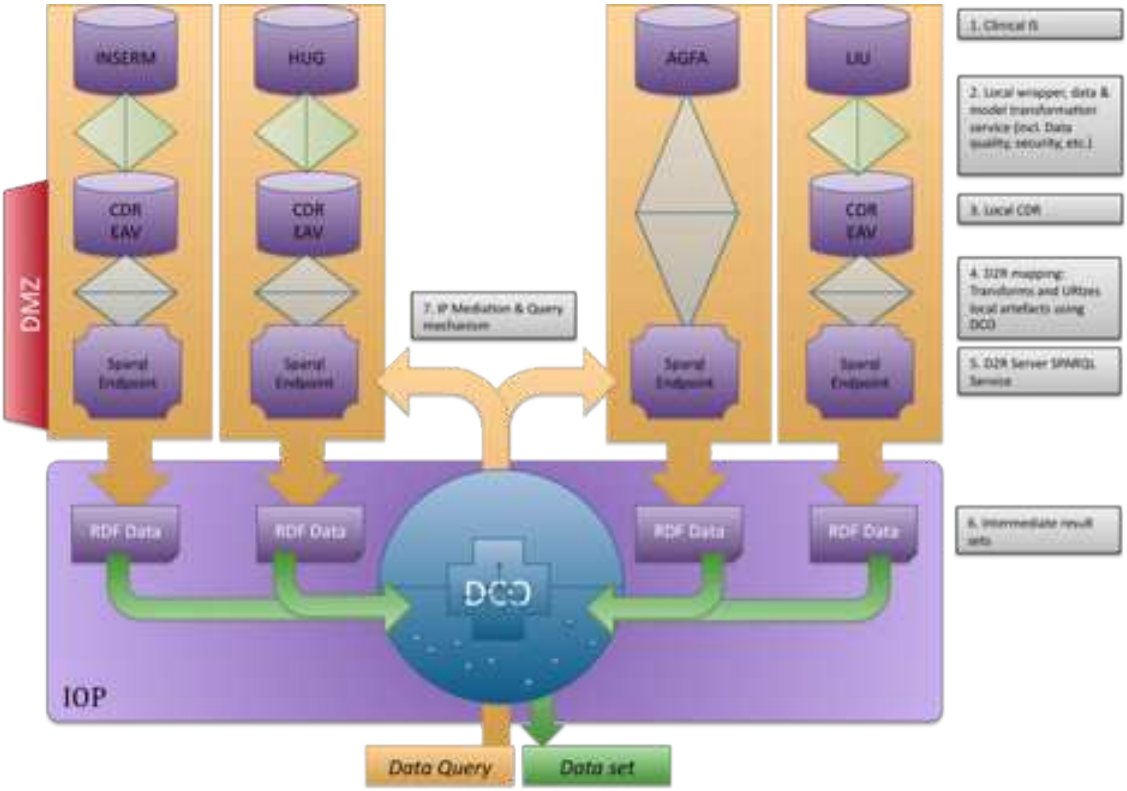


FIGURE 6.17 – Vue logique du système de médiation de données dans IP.

```
CONSTRUCT {dco:antibiotic a ?abg }
WHERE
{ ddo1:antibiotique a ?abg .}
```

Le résultat de cette requête si la clause WHERE est satisfaite, sera un ensemble de triplets décrivant des antibiotiques au sens DCO. On imagine alors comment une autre requête pour un deuxième CDR peut être :

```
CONSTRUCT {dco:antibiotic a ?abg }
WHERE
{ ddo2:drug a ?abg;
ddo2:typeDrug ?typ .}
FILTER
{ ?typ == 'antibiotic'}
```

Nous remarquons que les résultats des 2 requêtes seront équivalents. Par contre, la condition est très différente entre les 2 requêtes. Le graphe de sortie (DCO) de

la requête 1 est identique au graphe d'entrée (DDO1). Alors que le graphe de sortie de la requête 2 est différent du graphe d'entrée (DDO2). En effet, la DDO2 n'a pas de concept d'antibiotique, elle a par contre un concept de *drug* dans lequel un type est disponible afin de nous permettre de trouver les antibiotiques que nous cherchons. C'est afin de pallier ce type d'hétérogénéité dans les DDOs (CDRs) que nous proposons la mise en oeuvre de règles afin de relier les données de chaque CDR à la DCO.

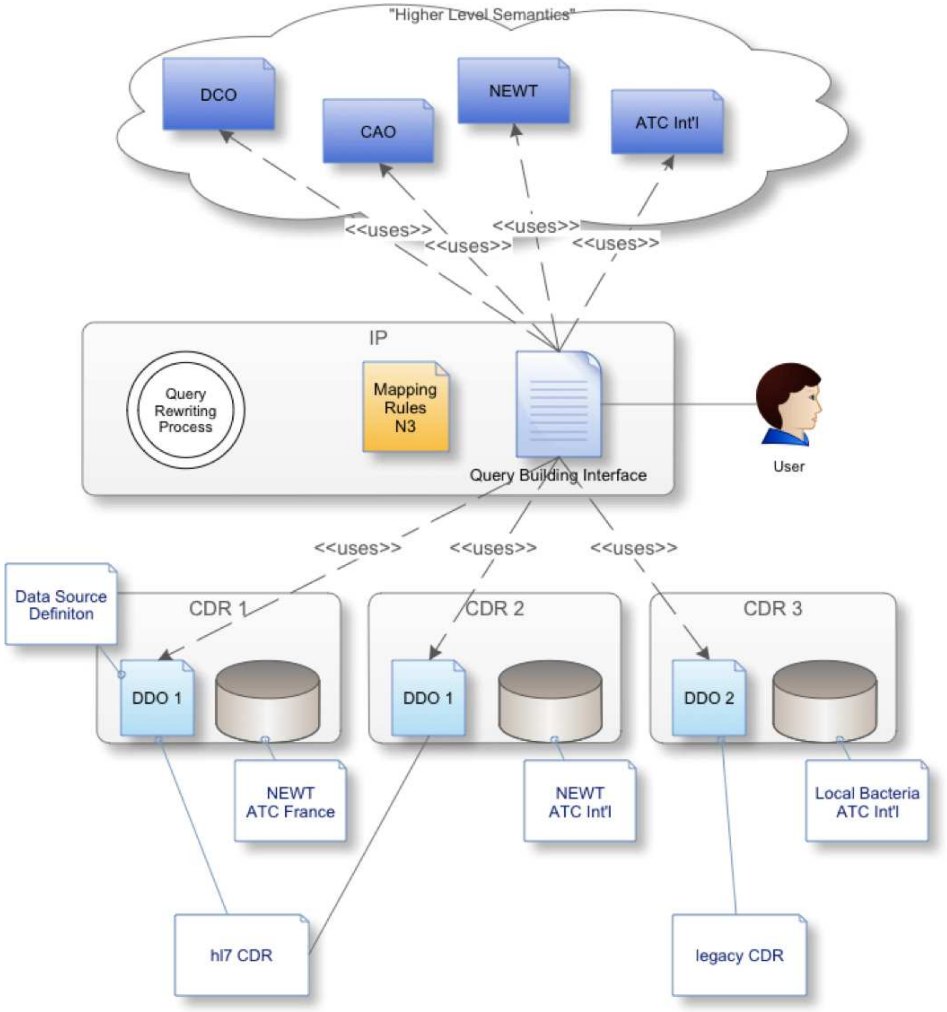


FIGURE 6.18 – Vue logique du système de médiation de données dans IP.

Les règles sont générées par les responsables de chaque "endpoint". En effet, ils détiennent le plus souvent la connaissance nécessaire à l'interprétation de leurs données. Voici un jeu de règles illustrant les 2 requêtes ci-dessus :

```
Rule 1:
```

```
{ ?x a dco:antibiotic } => { ?x a ddo1:antibiotique }

Rule 2:
{ ?x a ddo2:drug; ddo2:typeDrug ?type. ?type a ?C.
?C rdfs:subClassOf dco:antibiotic } => {dco:Antibiotic}

Rule 3:
{ ?x a dco:antibiotic } => { ?x a ddo2:Antibiotic}.
```

La première règle correspondant à la première requête ne pose pas de problème : *tout x étant un dco :antibiotic est un ddo1 :antibiotique*. La deuxième requête se décompose en 2 règles. Tout d'acord, *tout x étant un ddo2 :drug et ayant pour type une sous-classe de dco :antibiotic est un dco :antibiotic*. Puis, *pour tout x étant un dco :antibiotic, alors c'est un ddo2 :Antibiotic*.
Ces règles permettent de déduire certains éléments de la requête de données à effectuer sur les SPARQL endpoints. Il reste en effet à interpréter les relations DCO :concepts => DDO :instances ainsi que les éléments de vocabulaire du langage SPARQL.
Comme nous l'avons présenté plus haut, afin de relier les tuples de la base de données aux concepts de l'ontologie, nous définissons dans la DDO la portée des concepts correspondant aux valeurs possibles que peuvent prendre les concepts. Par exemple, pour la classe *Bacteria* :

```
ddo:Bacteria a rdfs:Class;
  rdfs:isDefinedBy ddo:Bacteria;
  rdfs:label """bacteria"""@en;
  skos:definition """An infectious agent..."""@en;
  rdfs:subClassOf [ a owl:Restriction;
                                    owl:onProperty ddo:bacteriaName;
                                    owl:someValuesFrom rdf:Literal ],
                                [ a owl:Restriction;
                                    owl:onProperty skos:annotation;
                                    owl:hasValue ts:newt ];
                                [ a owl:Restriction;
                                    owl:onProperty skos:inScheme;
                                    owl:someValuesFrom ts:newtDT ].
```

Nous déclarons ici que la classe Bacteria peut prendre comme valeur une valeur du vocabulaire newtDT représentant le datatype NEWT. Grâce à cette informa-

tion, le moteur de réécriture de requêtes saura faire le lien entre la bactérie issue de la DCO, où il sera annoté de son code NEWT. D'un point de vue de l'utilisation d'une ontologie de domaine pour effectuer des requêtes sur des sources distribuées, il faudra pouvoir gérer un cas de figure particulier. En effet, lorsqu'un utilisateur sélectionne un concept du type E.Coli au niveau de l'ontologie de domaine (servant de médiateur), nous n'avons pas de règle définie pour décrire la relation entre l'instance E.Coli du concept ddo :Drug et le concept dco :E.Coli.
Nous avons donc : E.Coli (un concept DCO) et Bacteria (un concept DDO) <– instanceOf – E. Coli (un tuple). Nous voulons pouvoir résoudre le problème en utilisant les propriétés de subsumption de l'ontologie de domaine. Nous savons, que ddo :bacteria => dco :bacteriaCell (nous rediscuterons du sens de la règle par la suite). Nous savons que dco :E.coli est une sous-classe de dco :bacteriaCell. Nous pouvons donc en déduire et généraliser que :

Definition 5. *si (prefixGlobal :Concept1) est sous-classe de (prefixGlobal :Concept2) et si (prefixLocal :Concept2) a pour conséquence logique (prefixGlobal :Concept2) alors (SELECT prefixGlobal :Concept1) a pour conséquence logique (SELECT prefixLocal :2' FILTER prefixGlobal :Concept2) où (prefixGlobal :Concept2) a pour annotation un code NEWT et (prefixLocal :Concept2) a pour datatype newtDT.*

Grâce à ces règles, nous pouvons résoudre le problème de réécriture de requête lorsque la modélisation des données est différente entre l'ontologie de domaine et les ontologies de données.

6.4.5 Le problème du monde ouvert

A la différence des systèmes d'intégration de données à la volée qui ont été présentés dans l'état de l'art, notre méthodologie tente de suivre autant que possible le paradigme du web sémantique : un monde ouvert. Un monde ouvert est un espace où l'inconnu peut exister. Dans une base de données relationnelle, lorsque l'on veut le nombre de personnes qui ne sont pas des femmes, une requête telle que celle-ci pourra être exécutée :

```
SELECT COUNT(*)
FROM personne
WHERE sexe (NOT IN 'F')
```

L'ensemble des valeurs possibles du champ personne est connu et fini. Lorsqu'on sélectionne un sous-ensemble de celui-ci et qu'on demande son inverse, le résultat est donné. Le problème du monde ouvert est que l'ensemble des valeurs que *personne*

peut prendre est potentiellement infini. C'est l'internet. Une nouvelle source de données peut apporter de nouvelles personnes à tout moment. L'inverse d'un sous-ensemble sur l'Internet est potentiellement gigantesque. Ce qui, à l'échelle, est un problème fondamental. En SPARQL, la même requête est résolue de la manière suivante :

```
SELECT COUNT(?personne)
WHERE
{
    ?personne foaf:hasSex _:sex .
    NOT EXISTS { ?personne foaf:hasSex 'F' . }
}
```

Cette requête est valide en SPARQL 1.1 [15]. C'est une requête qui a des solutions sur un SPARQL endpoint. Mais qui peut poser des problèmes sur une échelle plus grande, comme nous venons de le discuter. La fonction de négation est donc à utiliser avec précaution.

Au cours de notre expérimentation, nous avons été confrontés à un problème de généralisation de notre approche de réécriture de requêtes pour la médiation de données. A titre de rappel, le processus d'écriture tel que nous l'avions imaginé était le suivant :

- La requête est composée d'une partie CONSTRUCT en DCO(ressource médiatrice) et des règles sont utilisées pour créer les parties WHERE de chaque CDR,
- les résultats générés par les "endpoints" sont exprimés en DCO et ils n'ont plus qu'à être agrégés.

Une base de règles était créée avec des règles du type :

```
dco:A => ddo1:A
dco:A => ddo2:A
```

Ces 2 règles impliquent que :

```
dco:A => ddo1:A OU ddo2:A
```

.

Ce qui, dans un contexte de réécriture de requête donnerait la requête :

15. http://www.w3.org/TR/sparql11-query/ SPARQL 1.1 n'est pas encore implémenté au moment de l'écriture de ce manuscript

```
SELECT all dco:A
```

en 2 requêtes pour chaque source :

```
SELECT all ddo1:A
SELECT all ddo2:A
```

Mais cette règle implique aussi la règle suivante :

```
NOT ddo1:A => NOT dco:A
```

Ce qui est d'un point de vue logique, vrai. Mais faux d'un point de vue de la connaissance. Le fait que le concept A n'existe pas dans la source 1 n'implique pas qu'il n'existe pas dans l'ontologie de domaine. L'inverse par contre, dans le contexte de réécriture de requête est vrai :

```
NOT dco:A => NOT ddo1:A
```

Car si le concept A n'existe pas, alors la requête ne peut être posée puisque la requête s'exprime forcément dans les termes de l'ontologie de domaine. Bien entendu, le concept peut exister au niveau local sans forcément exister au niveau global, mais dans notre contexte d'utilisation, cela n'a pas d'impact.
Ensuite, même dans le cas de règles de réécriture d'un système unique, le problème existe, imaginons :

```
dco:A => ddo1:A OU ddo1:B
```

Cela créée une disjonction dans la conclusion (dco :A => ddo1 :A OU ddo2 :A) et va donc au delà de la logique de Horn qui est la portion décidable de la logique de premier ordre. Le système deviendrait non décidable. Enfin, si l'on rajoute une nouvelle règle à la base de règles précédente :

```
dco:A => ddo3:A
```

Alors la règle précédente sur dco :A n'est plus valide. Ce qui dans un contexte du web est un problème majeur. Ce qui est vrai hier doit rester vrai lorsqu'on rajoute des nouveaux systèmes.

6.4.6 Une proposition partiellement satisfaisante

Afin de réduire le problème précédemment abordé, nous proposons une méthodologie d'interopérabilité sémantique où nous mettons en oeuvre des règles bottom-up en lieu et place des règles top-down. Ces règles ddo(x) :Y => dco :Z ont la robustesse nécessaire à la mise en oeuvre d'un système qui doit pouvoir évoluer :

- Horizontalement : par l'ajout de nouvelles sources de données (ddo(1), ddo(2), ... , ddo(n)) et
- verticalement : par l'ajout de nouveaux concepts (ddo(x) :A, ddo(x) :B, ..., ddo(x) :N).

En effet, lorsqu'on ajoute une nouvelle règle à la base de règles, celle-ci soit mettra à jour la règle précédente, soit sera une nouvelle connaissance ajoutée. On dit alors que la base de règles est monotonique, les règles restent vraies quand une nouvelle règle est ajoutée. Considérons la base de règles suivante :

```
ddo1:A => dco:A
ddo2:A => dco:A
ddo1:B AND ddo1:C => dco:B
ddo2:B => dco:B
```

Si nous ajoutions une nouvelle règle, comme par exemple :

```
ddo2:B AND ddo2:C => dco:B, alors

(ddo2:B AND ddo2:C) OR ddo2:B => dco:B.
```

Ce qui est vrai. De même, si une règle évolue, sans qu'on ait besoin de supprimer la règle originelle (c'est à dire si elle était vraie à un moment donné), alors la prémisse(1) OU la prémisse(2) impliqueront le concept dco :x. Ce qui est aussi vrai.

Nous remarquerons cependant que si ce système de règles résout la deuxième partie de notre problème d'interopérabilité sémantique des données, elles permettent la réécriture des résultats, et non des requêtes. A ce jour, deux propositions ont été faites :

- Mettre en oeuvre des templates de questions pré-établies, où des 'variables' seront reliées aux concepts dco et,
- travailler sur une extension du moteur d'inférence afin de pouvoir gérer les cas logiques spécifiques à notre problème et ainsi permettre la réécriture top-down en se basant sur les règles bottom-up définies.

Nous évoquerons de ces approches dans la discussion de ce mémoire.

6.5 Conclusion

Nous avons abordé dans cette section divers aspects de l'interopérabilité sémantique. En premier lieu, la connaissance de la qualité de l'information que nous souhaitons partager est un atout pour évaluer la "marche sémantique" qu'il y a entre les données que nous voulons intégrer et le domaine dans lequel ces données seront utilisées. Nous pensons que cette étape d'évaluation de la qualité est nécessaire, nous proposons de gérer ces données qualité comme métadonnées et nous proposons une formalisation de celles-ci. De la même manière, pour partager de l'information, biomédicale ou non, l'expression du modèle de données et du vocabulaire dans lesquelles les données sont enregistrées est une composante essentielle. C'est dans ce contexte d'affichage de l'information à propos des données que nous proposons une formalisation d'une source de données. Cette ontologie de définition des données (Data Definition Ontology, DDO) peut être vue comme un modèle de données formalisé, où nous explicitons les concepts et relations. La DDO permet aussi, lorsque le modèle sous-jacent change, de ne pas changer la vue externe de la source de données. Les requêtes ne changent pas. L'intégration de notion de qualité de données dans la DDO permet d'offrir une vue multiple des données au monde extérieur. La figure 6.19 représente une vue du processus d'enrichissement et de formalisation d'une source de données telle que nous la proposons. Un premier processus de formalisation des données permet de décrire formellement les données du point de vue de la structure de stockage et de leur vocabulaire, ce processus est guidé par des experts. Le deuxième processus d'annotation de qualité permet, suivant les axes définis précédemment, d'enrichir l'ontologie de données de données relative à la qualité. La vue ontologique des données résultante à nos deux processus permet aux systèmes externes d'interroger les données via les concepts de l'ontologie, et de connaître le sens de ces données et leur qualité.

Dans un deuxième temps, nous avons proposé une plateforme d'interopérabilité permettant de prendre en compte, tout ou partie des propositions d'enrichissement des données pour le partage d'information. Cette plateforme d'interopérabilité à pour but premier de permettre d'échanger des données dans un domaine donné. Elle est tournée vers le monde ouvert du web sémantique et permet de prendre en compte la sémantique des données lors du processus de réécriture des résultats. La méthodologie proposée pour effectuer l'intégration de données sémantiques est, de notre point de vue, robuste pour le passage à l'échelle puisqu'elle est monotonique. C'est une approche bottom-up qui ne nécessite pas l'imposition d'une vue globale. Cependant, nous pensons que des données s'exploitent dans le cadre d'un domaine d'analyse particulier. C'est pourquoi nous avons mis en oeuvre une méthodologie

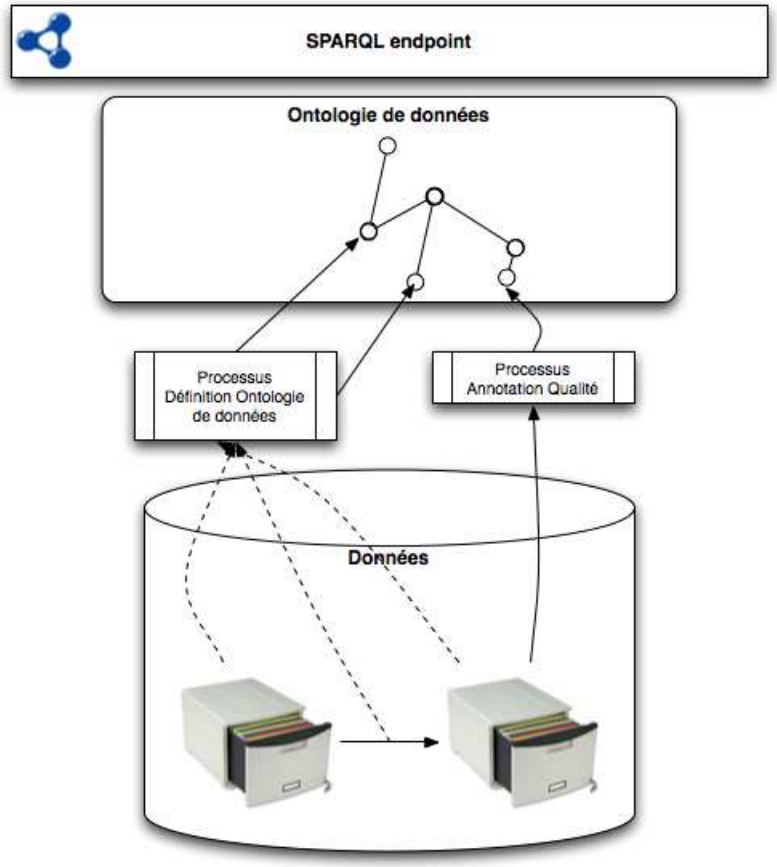


FIGURE 6.19 – Processus d'enrichissement et de formalisation d'une source de données pour le partage et l'analyse de données

d'alignement bottom-up. Cette stratégie d'alignement permet entre autres d'assurer :

- l'interprétation des données d'un point de vue du domaine, avec toute la robustesse logique d'une ontologie de domaine du point de vue de l'exploitation des données médicales,
- la liberté d'expression des données et des structures des sources de données (c'est un principe de réalité, les données sont imparfaites et imparfaitement stockées, car il n'existe pas de structure meilleure qu'une autre, elle est lié à l'usage),
- l'extension du système à d'autres sources et à d'autres domaines.

Notre approche a cependant des limites, principalement liées aux limites des systèmes à base de règles. Alors qu'il est plus adapté d'aligner du plus spécifique au plus général (ou du plus réel vers une organisation conceptuelle, une vue), il est

difficile d'utiliser la même base de règles pour faire de la réécriture automatique et de garder un système ouvert avec les raisonneurs dont nous disposons à ce jour. Nous avons, dans le cadre de notre expérimentation entrevu des options pour résoudre ce problème logique.

Chapitre 7

Expérimentation

"Expérimenter, c'est imaginer." - Frederich Nietzsche.

Sommaire

7.1 Introduction 148
7.2 Évaluation de la qualité 150
7.2.1 Mise en oeuvre 150
7.2.1.1 Audit 150
7.2.1.2 Qualification 151
7.2.1.3 Normalisation et surveillance 152
7.3 Constitution d'un entrepôt de données clinique 154
7.3.1 Entrepôt de données de santé : Une modélisation standardisée 154
7.3.2 Les autres Clinical Data Repository européens 155
7.4 La plateforme d'interopérabilité 157
7.4.1 Formalisation d'une source de données 157
7.4.2 La médiation sémantique de données 159
7.4.2.1 Architecture générale 159
7.4.2.2 Un exemple de médiation de données 163
7.4.2.3 Résultats 169
7.4.3 Validation de l'approche de médiation sémantique 170
7.5 Conclusion 172

Faire parler des données, voilà ce que nous essayons de faire. Nous avons bien créé des modèles, des référentiels, il n'en reste pas moins que seul l'homme peut comprendre ce qu'il a entré dans la machine. La communauté du web sémantique se demande comment cette masse de données peut être mieux exploitée par la machine. On commence à se dire qu'il faut formaliser le sens, la sémantique des données. C'est la sémantique qu'il faut relier, non les termes, non les structures de stockage. Dans ce chapitre nous allons tenter de donner du sens aux données et de les

mettre en commun dans le cadre de l'étude de l'évolution de la résistance aux antibiotiques en Europe.

7.1 Introduction

Le cadre d'expérimentation proposé par le projet DebugIT va nous permettre de valider nos hypothèses de recherche. A savoir :

- Comment utiliser des ontologies de domaine pour interroger et exploiter des données stockées dans des bases hétérogènes et non formelles ? (robustesse de l'inférence "métier", caractère implicite d'une base de données, médiation, montée en charge)
- Comment formaliser des bases de données biomédicales ? (modèles d'informations et ontologies, terminologies, qualité de données)
- Comment matérialiser la relation base de données - ontologie de domaine dans un environnement multi-sites ?
- Peut-on gérer le multilinguisme, plusieurs terminologies et plusieurs représentations de l'information à la volée au moment de l'interrogation ?
- Est-ce que le modèle RDF permet l'exploitation de jeux de données volumineuses dans un cadre d'analyse ?
- Peut-on utiliser le raisonnement pour inférer la relation données - ontologie de domaine ?
- Les technologies du web sémantique sont-elles adaptées au traitement d'informations issues du domaine biomédical ?

Nous aborderons ces questions à travers nos contributions exposées lors du chapitre précédent. Premièrement, lors du recueil de données, nous avons défini un modèle d'extraction et de qualité de l'information, ainsi qu'une méthodologie de mesure et d'amélioration de cette qualité, afin de proposer dans le cadre du projet des données exploitables. Nous proposons ensuite un modèle d'information extrait des modèles conceptuels standardisés d'HL7 où nous proposons une modélisation adaptée au requêtage de grands volumes de données sur différents centres de données et à l'utilisation de référentiels sémantiques. Enfin, nous proposons une plateforme d'interopérabilité utilisant la sémantique où les données seront liées à leur sémantique formelle telle qu'elle l'est décrite et partagée dans le projet. Nous évaluerons l'ensemble de nos propositions dans le cadre de cas d'utilisation dans le domaine de l'antibiorésistance.

Notre apport méthodologique et de mise en oeuvre présenté dans le chapitre

précédent se situe à différents niveaux de l'expérimentation. La figure 7.1 représente une vue générale détaillée du travail de sémantisation des données ainsi que notre contribution dans cette problématique générale.

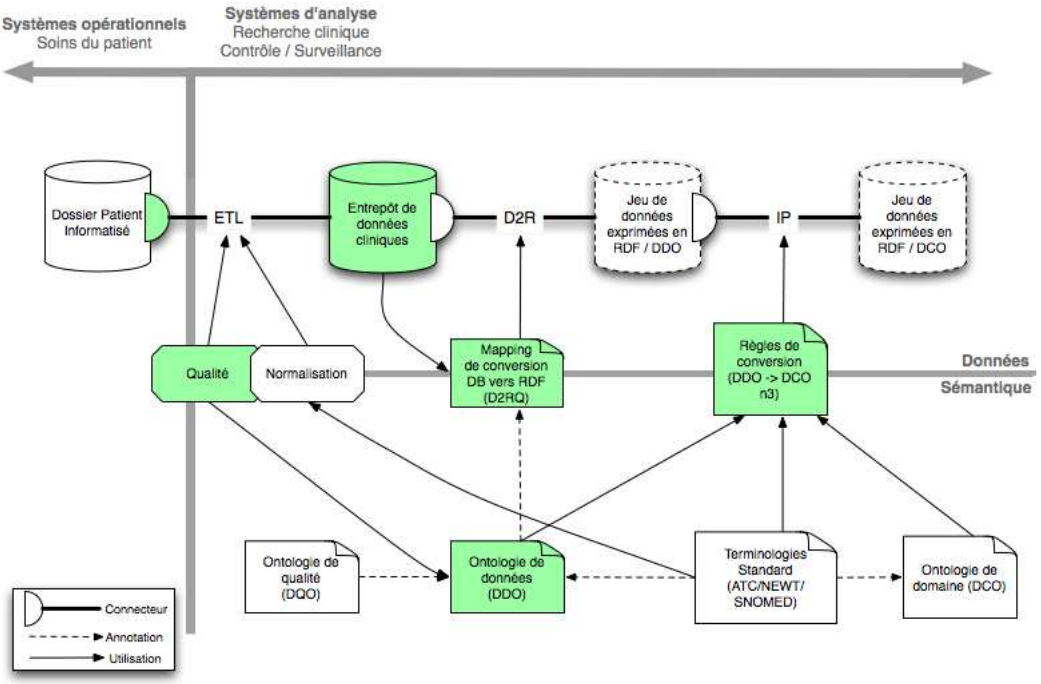


FIGURE 7.1 – Notre expérimentation se situera à différents niveaux dans la problématique d'interopérabilité de données (en vert)

Les données sont d'abord analysées en terme de qualité pour l'interopérabilité dans un cadre défini par le projet. Ensuite, nous expérimentons la mise en oeuvre d'un entrepôt de données clinique dimensionnel basé sur HL7. Nous présenterons ensuite des résultats de formalisation d'une ontologie de données. Enfin, nous présentons des règles de conversions locales pour achever le processus de sémantisation des données.

Dans un deuxième temps, nous présenterons des résultats d'agrégation de données sur différents sites dans le cadre de l'implémentation de la plate-forme d'interopérabilité.

Dans une dernière section, nous présenterons une validation de nos résultats par comparaison entre des rapports mis en oeuvre par des experts microbiologistes et les résultats obtenus grâce à la plateforme.

7.2 Évaluation de la qualité de l'information pour l'interopérabilité

Toute source de donnée n'est pas facilement utilisable et interprétable en l'état. Du moins, comme nous l'avons proposé dans les sections précédentes, il est nécessaire de mesurer et d'exposer la qualité d'une source d'information dans un contexte de partage de données. Nous proposons dans cette section de mettre en oeuvre notre méthode d'évaluation et de formalisation d'une source de données dans le cadre du projet DebugIT en mesurant la qualité de l'information source locale, le dossier patient informatisé de l'Hôpital Européen Georges Pompidou [1].

7.2.1 Mise en oeuvre

Nous allons d'abord définir un cadre de mise en oeuvre en utilisant la méthodologie en 4 étapes TDQM [2]. Pour chaque étape de la méthodologie TDQM, nous présentons ici les résultats de notre expérimentation pour les 3 sommets du triangle de la qualité de l'information.

7.2.1.1 Audit

Les objets sont mesurés à l'aide de procédures stockées sur la base de données source. Nous avons défini des procédures pour chaque champ que nous voulions tester. Les scores de qualité sont stockés dans une table indicateurs. La table 7.1 présente un extrait des résultats ainsi obtenus.

Objet	**Critère**	**Score**	**Commentaires**
DateFinSejour	Complétude	69,9%	Aide à calculer la durée du séjour
CodeUCD	Cohérence	75,6%	L'UCD est une classification française des noms de médicaments
PatientID	Unicité	100%	L'identifiant patient doit être unique

TABLE 7.1 – Extrait des scores relatifs aux objets

1. L'Hôpital Européen Georges Pompidou comprend 814 lits et gère 230000 consultations par an et 54000 admissions. Cet hôpital gère d'une part une mission de proximité en réponse aux besoins de santé de la population de l'Ouest Parisien et assure d'autre part des soins d'hospitalisation aigüe dans les pôles d'activité urgences-réseaux, cancérologie et spécialtés médicales et enfin en cardio-vasculaire. 8 unités INSERM lui sont rattachés ainsi que 2 unités CNRS. Enfin, un centre d'investigation clinique, une unité de recherche clinique et une unité épidémiologique sont présents.

2. Total Data Quality Management

Concernant les termes, la table 7.2 montre un exemple de mesure statistique de distance entre le référentiel des données, et le référentiel NEWT, ainsi que 2 référentiels locaux construits par un expert concernant les localisations et les types de prélèvement biologiques. La distance ainsi mesurée représente le pourcentage de termes alignés avec la ou les terminologies de référence grâce aux scripts PERL développés aux hôpitaux universitaires de Genève.

Terminologies (de référence)	**Distance au standard**
Noms des bactéries / NEWT	85,53%
Antibiotiques Prescrits / ATC	67%
Localisation de prélèvement / SNOMED	46%

TABLE 7.2 – Distance statistique aux référentiels

Enfin, concernant les concepts ou la modélisation conceptuelle du domaine d'intérêt de notre étude, nous avons appliqué la méthode d'évaluation subjective des modèles d'information. La figure 7.2 présente le résultat obtenu sur le modèle d'information du DPI de l'HEGP, avec le modèle d'information HL7 comme référence. Chaque axe est mesuré de façon empirique (avec l'aide d'un expert du domaine), et noté de 1 (mauvais) à 5 (excellent). Beaucoup d'exemples peuvent illustrer la distance entre le modèle du DPI et celui d'HL7. Par exemple, dans le domaine des résultats de laboratoire, le modèle « Result Event » d'HL7 propose une classe spécifique pour gérer les groupes de tests à faire sur un spécimen dérivé d'une culture donnée, ce qui permet une gestion plus fine des résultats microbiologiques comparé à celle en vigueur à l'HEGP à ce jour. Du point de vue de la compréhension du modèle d'information, C_SPECIALITE, qui en fait est une table contenant les prescriptions médicamenteuses par spécialité, n'est pas un terme parlant. D'un autre côté, une force du modèle d'information du DPI est qu'il intègre un référentiel d'éléments de données partagés, qui permet de lier le référentiel du DPI à d'autres référentiels de termes, ceci aidant à l'intégration.

7.2.1.2 Qualification

La qualification de la source d'information pour le domaine restreint de notre étude est présentée dans la table 7.3.

Cette phase de qualification peut être considérée comme une validation de notre TQI. Le résultat de l'évaluation de la dimension conceptuelle du DPI évalué était 2.42, son score qualité sera donc C. Le score global reflète la qualité du DPI étu-

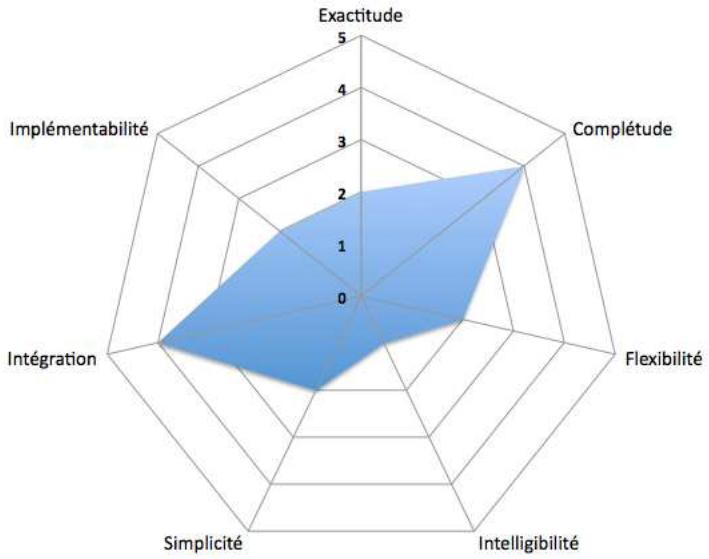


FIGURE 7.2 – Résultat de la méthode d'évaluation subjective du modèle d'information du DPI de l'HEGP

Sommet	**Score**
Objet	C
Terme	B
Concept	C
Global	C

TABLE 7.3 – Evaluation de la source de données DPI de l'étude

dié dans un cadre d'interopérabilité dans le domaine des maladies infectieuses. Il représente aussi la distance entre l'information source et le domaine sémantique.

7.2.1.3 Normalisation et surveillance

Le processus de normalisation a été effectué lors du chargement des données vers notre entrepôt de données de santé pour la recherche clinique. Tout d'abord, nous avons extrait les référentiels locaux du DPI de l'HEGP dans le cadre de notre expérimentation. Nous avons identifié les référentiels suivants, susceptibles d'être standardisés :

- Les antibiotiques prescrits,
- les bactéries testées,
- les substances,

– les services hospitaliers,
– les localisations de prélèvement.

Chaque concept est susceptible d'être interrogé à un niveau global. Nous avons donc mis en oeuvre dans le cadre du projet, avec l'aide d'outils d'alignement divers (techniques de traitement automatique du langage, techniques de mapping), une normalisation de nos données. Nous avons utilisé des référentiels contrôlés et standards. Puis, nous avons mesuré le taux de couverture de normalisation. Dans certains cas le taux de normalisation en multi-annotation (plusieurs codes pour le même tuple). Enfin, le taux d'annotations trouvées. Le tableau suivant résume les résultats obtenus :

Référentiel	**Domaine**	**Couverture**	**Multi-Annotation**	**Nb Annotations/ Nb Termes**
ATC	Antibiotiques prescrits	92%	51%	22388/8170
ATC	Substances testées	87%	0%	62/71
NEWT	Bactéries testées	90%	0%	157/175
SNOMED	Localisations	86%	0%	81/94

Table 7.4 – Taux de normalisation des termes par terminologie standard après normalisation automatisée

Nous avons rencontrés plusieurs difficultés lors de la phase de normalisation. Tout d'abord, chaque terminologie n'est pas forcément adaptée à l'annotation de données. L'ATC, comme nous l'avons vu dans l'état de l'art, a été construite pour des usages cliniques particuliers. Elle présente une classification multi-niveaux ou le même concept (médicament) peut être classé à différents endroits (doublons). Sachant que nous voulons annoter des médicaments donnés dans un cadre général de prescription, et, des substances testées dans un cadre d'analyse microbiologique, l'ATC est très peu adaptée à l'usage que nous voulons en faire, à savoir simplement annoter des données. De plus, l'ATC recense des substances actives qui sont parfois des noms de médicaments, et parfois contenues en combinaison dans un médicament avec une autre substance active. Par exemple, prenons l'amoxicilline. Voici la liste des codes contenant de l'amoxiciline dans l'ATC : J01CA04, J01CR02, A02BD06, A02BD07, A02BD03, J01DD06, A02BD05, A02BD01, A02BD04. Parfois en combinaison avec un autre produit, parfois non. Dans tous les cas, lorsqu'on cherche les patients à qui on a donné de l'amoxicilline, on doit permettre la récupération des instances. Puisqu'il n'y à pas de propriété d'équivalence dans l'ATC (ce n'est pas

une ontologie), il nous faut annoter chaque instance de la base de données avec tous les codes ATC possibles.
Par ailleurs, nous avons mis en oeuvre des procédures de normalisation automatiques qui incluent un système de détection de nouveaux concepts afin de pouvoir enrichir la normalisation manuellement le cas échéant.

7.3 Constitution d'un entrepôt de données clinique

Au démarrage du projet DebugIT, plusieurs partenaires pensaient pouvoir connecter au réseau de données DebugIT les dossiers patients informatisés directement via des connecteurs dynamiques ou des vues. Après une analyse complète de la qualité de l'information du DPI de l'HEGP ainsi que de son modèle de données, nous nous sommes ravisés. En effet, un DPI est un outil de capture et de gestion de l'information clinique au jour le jour. Il est d'une part construit dans un but de mise à jour de l'information et d'autre donne une liberté au professionnel de santé dans l'entrée de l'information. Enfin, la sécurité des systèmes d'information intra-hospitaliers ne permettent pas de connexion directe au DPI, pour des raisons de performances et de sécurité. Il est donc assez classique de créer une couche de données matérialisées dans un entrepôt de données.
De plus, les modèles relationnels adaptés à l'OLTP (on-line transactional processing) ont des performances beaucoup moins bonnes que les modèles dimensionnels adaptés à l'interrogation de grand volumes de données (voir Chapitre 2). C'est dans ce contexte que nous avons proposé l'implémentation d'un entrepôt de données clinique, dimensionnel et standard à HL7. Nous présentons dans cette partie la mise en oeuvre de notre entrepôt.

7.3.1 Entrepôt de données de santé : Une modélisation standardisée

Comme nous l'avons proposé dans la section 6.3.2.1, nous proposons la mise en oeuvre d'une modélisation dimensionnelle standardisée au domaine clinique de notre étude. Le modèle résultant se décompose en :

- une table de faits SubstanceAdministrationRequest visant à enregistrer l'information relative à la prescription de médicaments avec une granularité des tuples (de chaque enregistrement) au niveau de l'acte de prescription, et non de la prise réelle du médicament qui est une information de nature différente, la mesure de ce fait peut être la quantité de prescription, la fréquence d'administration,

- une table de faits AntibiogramObservationEvent représentant un test microbiologique et ayant pour mesure le résultat de ce test,
- plusieurs tables de dimensions communes aux deux tables de faits : StepOfStay (étape dans le parcours de soin du séjour) qui a pour hiérarchie de dimension Stay (le séjour), et qui a pour tables relationnelles (non dimensionnelles) l'AdministrativeDiagnosis (le diagnostic) et les Procedures (les actes),
- la table Patient, AssignedEntityOrganization (représentant le service hospitalier), deux dimensions temps EffectiveTimeYear et EffectiveTimeMonth,
- le groupe de tables SpecializedKind-ATC représentant le codage ATC correspondant aux médicaments et substances (Medicine) prescrits.
- enfin, des dimensions propres à la table de faits AntibiogramObservationEvent comme CultureObservationEvent qui représente la notion de culture microbiologique, Specimen pour les notions de localisation du prélèvement, BacterialColony pour la notion de colonie bactérienne et enfin ObservationReport pour la notion de rapport microbiologique.

La principale difficulté de modélisation était d'être capable de respecter la composante dimensionnelle du modèle au niveau des médicaments et des substances. En effet, un médicament peut contenir plusieurs substances actives. Une substance testée en microbiologie peut être associée à une autre substance. Et nous voulons bien évidemment annoter avec l'ATC d'une part les médicaments prescrits et d'autre part les substances testées. La modélisation dimensionnelle recommande de garder un lien de spécialisation -> généralisation de la table de faits vers les dimension et la hiérarchie de dimension. Le grain le plus fin serait donc la substance. Il faudrait cependant ramener au niveau de la substance les médicaments prescrits afin de pouvoir croiser les informations et définir une dimension commune. Ne possédant pas cette information, nous avons choisi d'utiliser l'ATC comme vecteur de liaison, en créant artificiellement une annotation simple sur les médicaments ou substances dans la classe des J01 (antibiotiques). Nous discuterons de ce choix dans la validation des résultats.

La figure 7.3 représente une version graphique du modèle dimensionnel de notre étude. Le chargement de l'entrepôt de données s'est effectué à l'aide d'outils ETL (Talend OpenStudio).

7.3.2 Les autres Clinical Data Repository européens

Chaque CDR européen a abordé sa propre méthode de constitution des données dans le cadre de DebugIT. Diverses techniques de modélisations ont étés mis en oeuvre. Aux Hôpitaux Universitaires de Genève**??**, a été constitué un pré-entrepôt

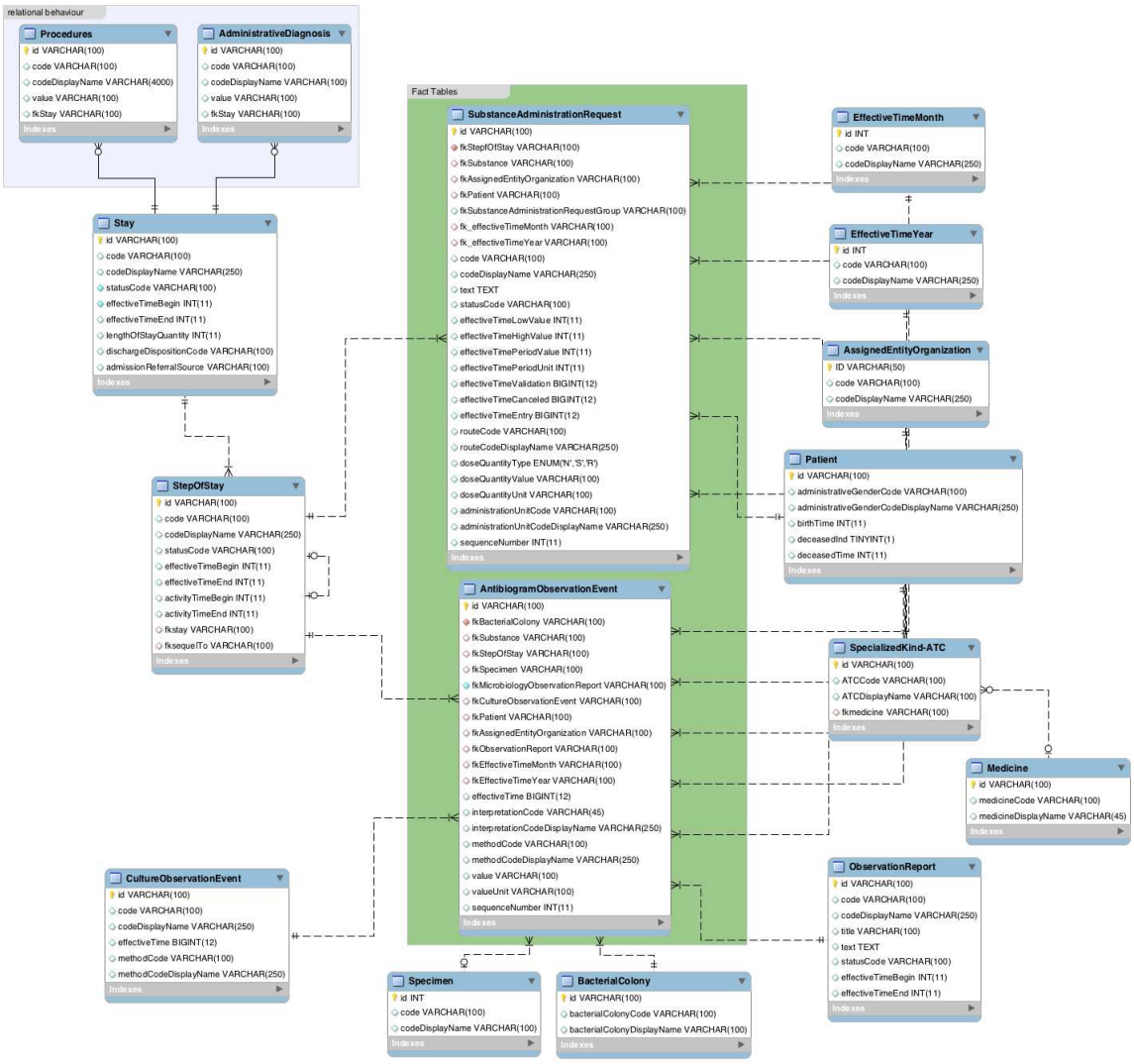


FIGURE 7.3 – Modèle dimensionnel HL7 pour l'étude de l'évolution de la résistance aux antibiotiques dans DebugIT

EAV afin d'aider à l'intégration de nouvelles données au fil du projet. Aux hopitaux de Linkoping en Suède, openEHR a été utilisé. En Allemagne, une autre utilisation d'HL7 a été faite. Dans tous les cas, la plateforme d'interopérabilité, malgré l'utilisation d'un standard, se doit d'offrir les outils pour intégrer ces différentes vue des mêmes données d'un domaine d'étude.

7.4 La plateforme d'interopérabilité

Comme nous l'avons vu dans la section précédente de notre expérimentation, les données disponibles pour notre étude sont exprimées dans des modèles différents, et avec des vocabulaires potentiellement différents. Cependant, dans le cadre du projet DebugIT, rappelons que diverses contraintes sont définies :

- Une vue unique et homogène des données du projet DebugIT doit être mise en œuvre.
- L'accès aux données doit être transparent, tant au niveau technique, syntaxique que sémantique.
- Les contraintes de confidentialité associées aux données ne nous permettent pas de stocker les données dans un entrepôt de données de manière centralisée. Chaque pays a d'ailleurs ses propres législations concernant la politique de confidentialité des données, et la méthode proposée devra pouvoir en tenir compte.
- L'accès direct aux données des dossiers patients depuis l'extérieur est donc généralement impossible pour des raisons de sécurité.

Par ailleurs, l'utilisateur devra avoir accès aux données grâce à une vue unique, robuste et formelle de son domaine : la core ontologie de DebugIT. Cette ontologie doit permettre à l'utilisateur d'interroger les données issues des CDR hétérogènes et doit représenter formellement le domaine étudié d'un point de vue médical. Nous présentons dans la section suivante les résultats de formalisation d'une source de données et de son expression en RDF. Ensuite, nous aborderons la partie médiation sémantique de la plateforme d'interopérabilité. Enfin, nous présenterons notre approche de validation de notre expérimentation.

7.4.1 Formalisation d'une source de données

Nous avons appliqué à notre source de données la méthode proposée dans le chapitre 6 afin de générer une vue RDF des données modélisées. Une ontologie de données a été construite grâce à un processus semi-automatisé s'appuyant sur le processus de génération de fichier D2R fourni avec la distribution de D2R server [3]. Le fichier résultant est ensuite revu par un expert des données afin de définir les vocabulaires adéquats aux concepts (dans notre expérimentation, les antibiotiques, les noms de bactéries, ou les localisations de prélèvement). La vue ainsi créée est utilisée dans le fichier D2R mapping qui sera lui-même utilisé par D2R pour générer les données en RDF annotées des bons concepts. Prenons l'exemple de la table

3. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/)

BacterialColony de notre modèle dimensionnel HL7.

```
map:Bacteria a d2rq:ClassMap;
  d2rq:dataStorage map:debugit;
  d2rq:class ddo:Bacteria;
  d2rq:uriPattern "BacterialColony/@@BacterialColony.id@@";
  .

[] a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Bacteria;
  d2rq:property skos:notation;
  d2rq:column "BacterialColony.bacterialColonyCode";
  d2rq:datatype biosko:uniProtTaxonomyDT;
  .

[] a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Bacteria;
  d2rq:property rdfs:label;
  d2rq:column "BacterialColony.bacterialColonyCode";
  .
```

Il est à noter que la DDO utilisée dans notre exemple est une DDO correspondant à une version précédente de notre entrepôt de données non HL7 et non dimensionnel. Nous continuons à l'utiliser ici car elle nous permet de ne pas changer la base de règles servant à lier les données à l'ontologie de domaine que nous verrons plus loin. Voici un extrait de la DDO correspondant au concept de bactérie.

```
ddo:Bacteria # maps to biotop:BacterialColony and cao:Bacterium
  a rdfs:Class;
  rdfs:isDefinedBy <http://debugit1.spim.jussieu.fr/ddo>;
  rdfs:label """bacteria"""@en;
  skos:definition """Aggregation of bacteria growing together as
  offspring of an unicellular organism without intracellular
  membranes, different from Archaeum mainly in DNA
  replication."""@en;
  rdfs:subClassOf [ a owl:Restriction;
    owl:onProperty ddo:hasBacteria_name;
    owl:someValuesFrom ddo:Bacteria_name],
```

```
[ a owl:Restriction;
  owl:onProperty skos:notation;
  owl:someValuesFrom biosko:uniProtTaxonomyDT ],
[ a owl:Restriction;       # OPTIONAL
  owl:onProperty skos:inScheme;
  owl:hasValue biosko:uniProtTaxonomy ].
```

La DDO représente une vue 'formelle' ou 'conceptuelle' du modèle de données de la base. La version complète de la DDO formalisée est disponible ici : https ://debugit.spim.jussieu.fr/ddo. Elle a la particularité de représenter les valeurs de vocabulaire possibles dans un champ. C'est une méthode que nous avons trouvé afin de pallier la problématique d'annotation des tuples de la base de données. En effet, un vocabulaire contrôlé comme l'ATC représente potentiellement un grand nombre de termes, et donc, même dans le cas où la base est normalisée, le système de médiation en amont doit connaître le nom de ce vocabulaire de termes afin d'en déduire le mécanisme de réécriture approprié. Le vocabulaire utilisé est aussi utilisé pour annoter les concepts dans l'ontologie de domaine.

7.4.2 La médiation sémantique de données

Forts de la formalisation de nos données effectuée dans la phase précédente de notre expérimentation, nous pouvons maintenant mettre en oeuvre la plateforme d'interopérabilité pour la médiation sémantique de données. Comme nous l'avons évoqué, les systèmes de gestion de données source seront médiées suivant l'approche GLAV où l'ontologie de domaine est la ressource médiatrice et où les ontologies de données locales sont rapprochées de l'ontologie de domaine par le biais de règles DDO -> DCO. Nous décrivons dans cette section tout d'abord l'architecture générale de l'IP qui représente le résultat des spécifications fonctionnelles décrites dans le chapitre 6. Ensuite, nous présenterons le processus de réécriture suivant un exemple, et nous présenterons nos résultats à l'issue de ce processus.

7.4.2.1 Architecture générale

Les spécifications fonctionnelles de la plateforme d'interopérabilité détaillées dans le chapitre 6 nous permettent de définir une architecture orientée services de la plateforme[Choquet 2011]. Tout d'abord, la méta-architecture est la suivante :

– Le web sémantique apporte un framework qui permet l'échange de données. Basé sur RDF, le framework permet le raisonnement. Le web sémantique ap-

porte une distinction entre le "monde" et une théorie sémantique formelle à propos de ce "monde".
- L'utilité d'une théorie sémantique formelle est qu'elle offre une solution technique pour déterminer si un raisonnement est vrai.
- Le moteur de réécriture à base de règles permet l'utilisation de règles pour effectuer de la médiation sémantique entre différents fournisseurs de données.
- Des web services sémantiques sont utilisés pour définir l'architecture d'accès aux ressources (données, requêtes, règles, terminologies, etc.).
- Les interfaces utilisateur s'appuient sur l'architecture orientée web services sémantiques pour avoir accès aux ressources du projet.

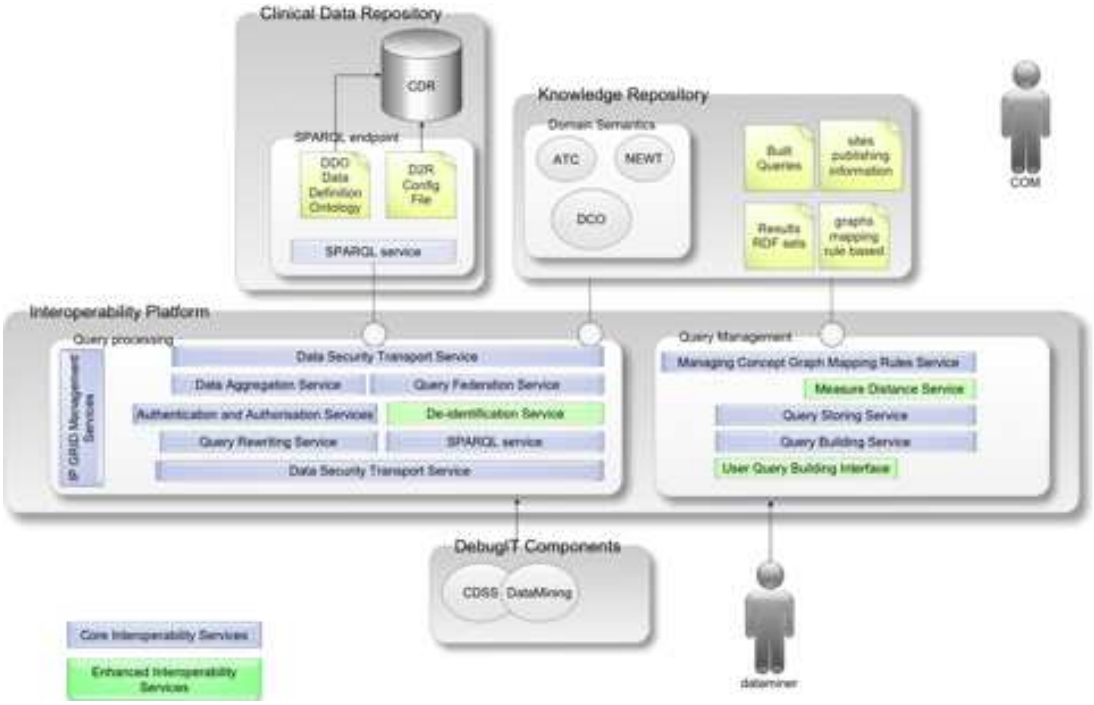


FIGURE 7.4 – Vue conceptuelle de l'architecture générale d'IP, de ses services et de ses acteurs.

La figure 7.4 présente une vue conceptuelle de la plate-forme d'interopérabilité, de ses services et de ses modules divisés en deux familles : le management des requêtes (Query Management) servant à la gestion des requêtes et à l'interface utilisateur ; et l'exécution des requêtes (Query Processing) regroupant les services de médiation sémantique :

- Le service Data Aggregation Service permet l'agrégation de données exprimées dans des formalismes différents suivant un formalisme partagé,

– le service Query Rewriting Service permet la réécriture de requêtes à partir de règles,
– le service Authentification and Autorisation Services assure la sécurité d'accès aux données,
– le SPARQL Service permet l'accès aux services de la plateforme,
– le Data Security Transport Serivce permet le transport sécurisé des données.

Nous définissons ensuite des packages suivant le formalisme UML (7.5)pour décrire les interactions possibles entre les services de la plateforme.

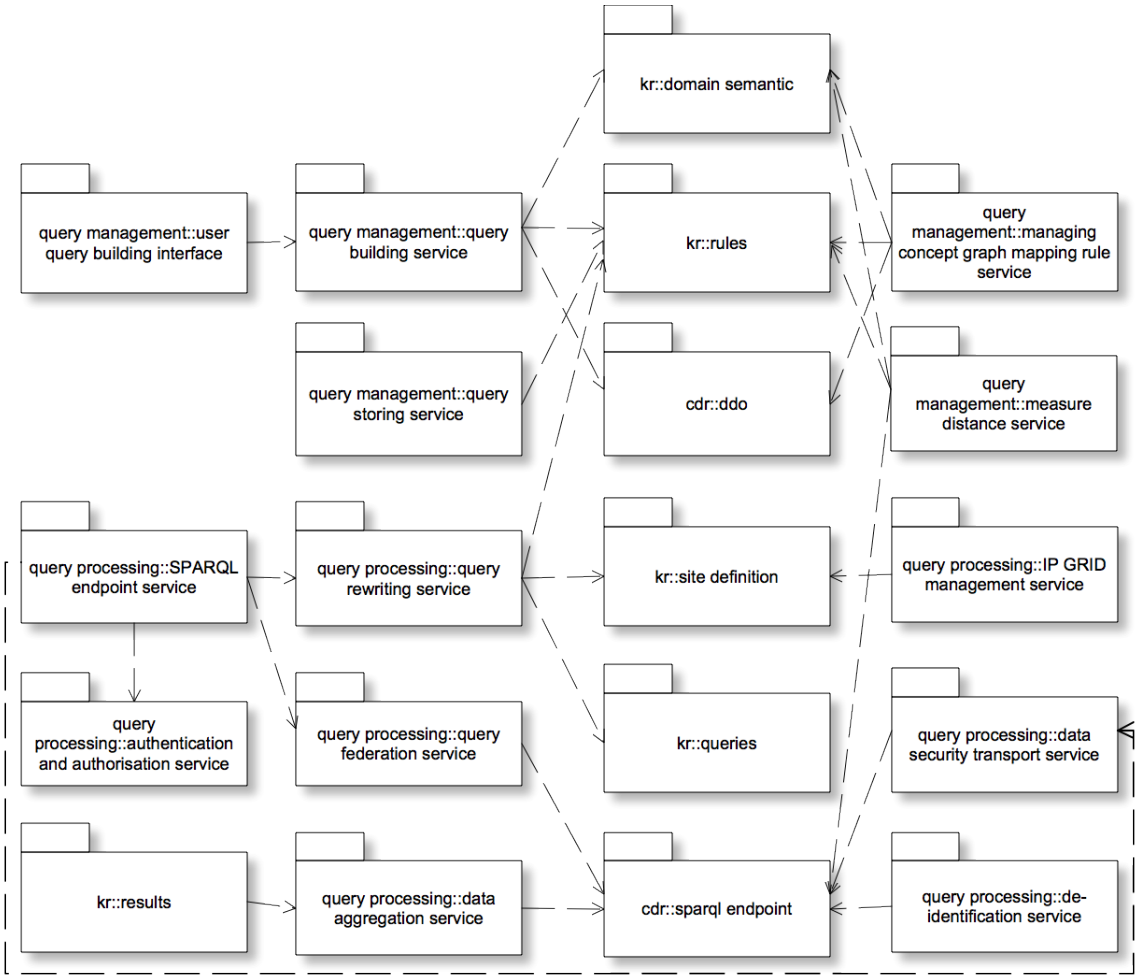


FIGURE 7.5 – Vue globale des services de IP et de leur interactions

Nous détaillons ensuite la mise en oeuvre des uses cases par l'élaboration de diagrammes de séquence. La figure 7.6 représente un diagramme de séquences de réécriture de requête basé sur des règles de réécriture stockées dans le "Knowledge

Repository".

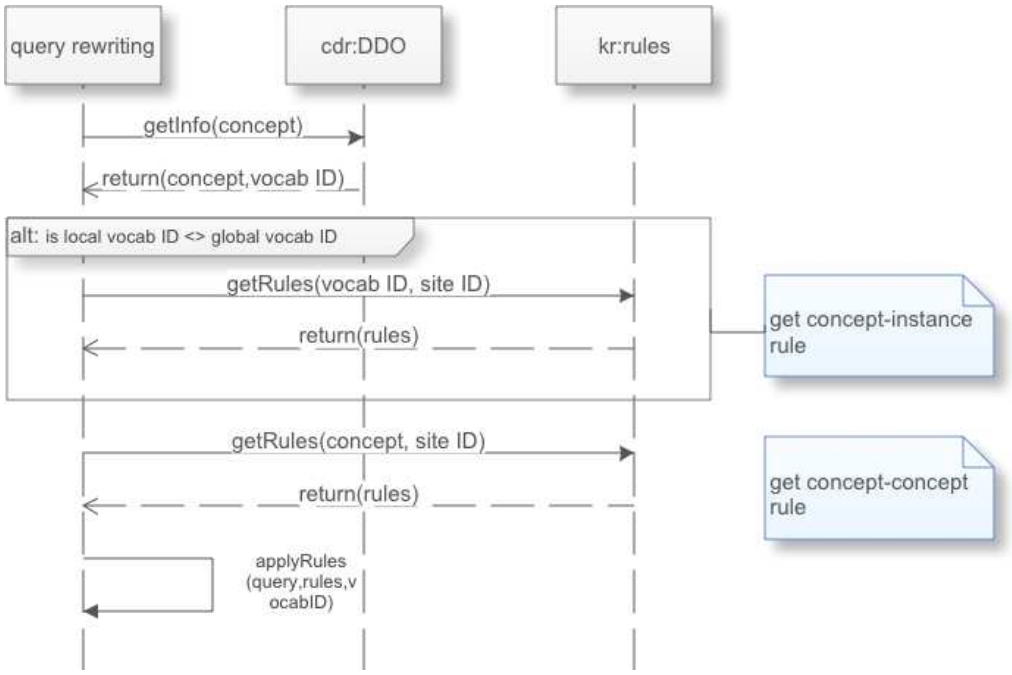


FIGURE 7.6 – Diagramme de séquence de réécriture de requête.

La totalité des diagrammes correspondant à l'architecture de IP sont disponibles dans le livrable D1.3 du projet DebugIT accessible à l'adresse suivante : `http://www.debugit.eu/progress/reports.html`

L'architecture proposée a été implémentée comme telle dans la majeure partie des cas. Le service Query Rewriting a cependant été revu suite aux limitations que nous avons présenté dans le chapitre 6. Notre première solution consiste en la création manuelle des requêtes pour chaque CDR en fonction d'une requête globale créée par l'utilisateur. La deuxième solution consiste en la création de patrons de requêtes que l'on formalise afin de permettre la réécriture partielle des requêtes suivant ceux-ci.

La figure 7.7 présente un diagramme de création d'une nouvelle requête clinique (exprimée en DCO, voir la section suivante) ainsi que les différentes étapes nécessaires afin de pouvoir exprimer ces résultats dans le même formalisme. Nous présentons aussi ici les différents acteurs ainsi que leurs tâches respectives afin de permettre à la plateforme d'exécuter les requêtes.

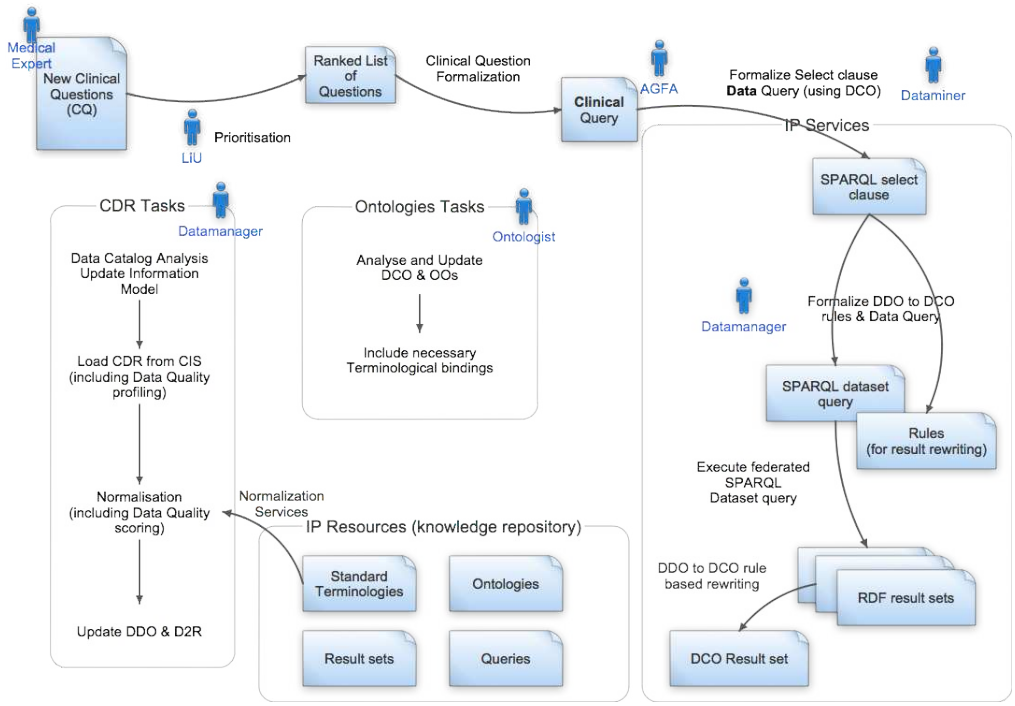


FIGURE 7.7 – Diagramme représentant le flux de tâches nécessaires à la création d'une nouvelle requête clinique dans DebugIT.

7.4.2.2 Un exemple de médiation de données

Afin d'illustrer la mise en oeuvre de la plateforme d'interopérabilité, nous allons présenter dans cette section un exemple d'exécution d'une question clinique sur différents CDR européens. Pour nous aider dans les différents processus ou services impliqués dans l'exécution d'une requête, la figure 7.8 représente un flux d'exécution de requête sur 2 CDRs.

L'utilisateur formalise une requête clinique en utilisant les concepts de la DCO. Par exemple : "What is the percentage of <bacteria> resistant to <antibiotic drug> in <sample-type> during <period=from start to current- date> from <site=here> ? "

Le patron définit une requête type avec des éléments variables entre < >. L'utilisateur définit une requête : "What is the percentage of S. Aureus cases, cultured from all samples, collected by a all sample collections, that is resistant to vancomycin, in the period from 1 January 2007 to (not including) 1 January 2008 at HEGP hospital ?"

La formalisation de cette requête en SPARQL créée au moment de l'exécution de la requête grâce aux patrons est de la forme :

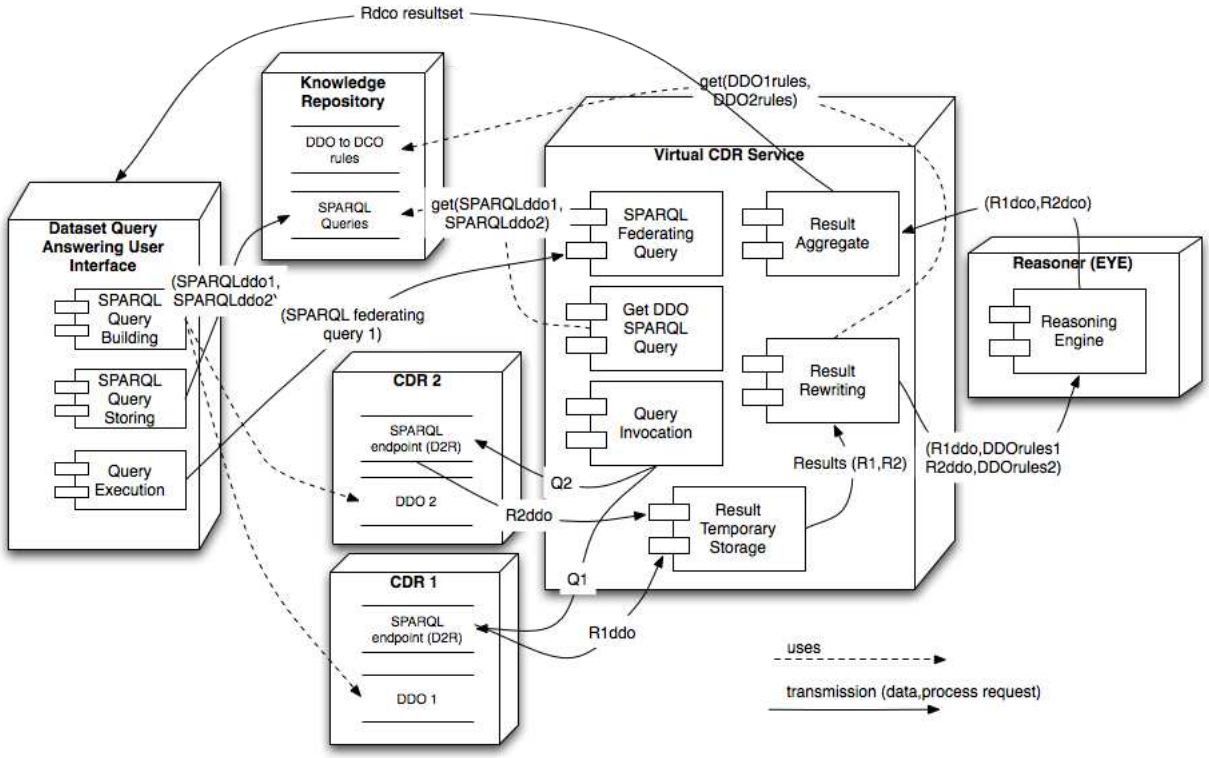


FIGURE 7.8 – Diagramme d'exécution d'une requête clinique sur 2 CDRs

```
WHERE {
 _:percentage
  quant:percentageOf _:total; # set
  quant:percentageThat _:part;
  quant:hasMeasurement [quant:hasValue ?percentageValue;
    quant:hasFactor quant:percent].
 _:total
  quant:counts ?total; # number
  rdfs:subClassOf cao:SAureus, [
   a owl:Restriction; owl:onProperty cao:culturedFrom;
    owl:someValuesFrom [
    rdfs:subClassOf dco:Sample, [
     a owl:Restriction; owl:onProperty biotop:outcomeOf;
      owl:someValuesFrom [
      rdfs:subClassOf dco:SampleCollection, [
```

```
       a owl:Restriction; owl:onProperty event:during;
        owl:someValuesFrom [
        dco:hasStartDateTime "2007-01-01T00:00:00"^^xsd:dataTime;
        dco:hasEndDateTime "2008-01-01T00:00:00"^^xsd:dataTime]], [
         a owl:Restriction; owl:onProperty biotop:hasLocus;
         owl:someValuesFrom dco:Hospital]]]]].
_:part
 quant:counts ?part;
 rdfs:subClassOf _:total, [
  a owl:Restriction; owl:onProperty cao:resistantTo; owl:someValuesFrom
   dco:Vancomycin]}
```

Nous remarquons dans cette requête que plusieurs ressources sont utilisées pour la formalisation des concepts et des relations. Ces ressources sont définies par les préfix suivants :

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX event: <http://eulersharp.sourceforge.net/2003/03swap/event#>
PREFIX quant: <http://eulersharp.sourceforge.net/2003/03swap/quantities#>
PREFIX units: <http://eulersharp.sourceforge.net/2003/03swap/units#>
PREFIX biotop: <http://purl.org/biotop/biotop.owl#>
PREFIX dco: <http://purl.org/imbi/dco/dco#>
PREFIx cao: <http://www.agfa.com/w3c/2009/clinicalAnalysisOntology#>
```

Suivant les règles de réécriture à partir des templates définies, la requête est réécrite pour chaque CDR/DDO. Voici une règle de réécriture servant à déduire le code UniPROT (NEWT) à partir d'une bactérie :

```
{
    [
        r:variable [ n3:uri "http://localhost/var#bacterium"];
        r:boundTo ?bacterium
    ] a r:Binding.
       ?bacterium skos:exactMatch [ skos:inScheme biosko:uniProtTaxonomy;
```

```
         skos:notation ?uniProtCode].
} =>{
        [
         r:variable [ n3:uri "http://localhost/var#uniProtCode"];
         r:boundTo ?uniProtCode
    ] a r:Binding.
}.
```

La requête obtenue pour l'HEGP est la suivante :

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX ddo: <http://debugit.spim.jussieu.fr/ddo#>
PREFIX biosko: <http://eulersharp.sourceforge.net/
2003/03swap/bioSKOSSchemes#>
PREFIX clisko: <http://eulersharp.sourceforge.net/2003/
03swap/clinicalSKOSSchemes#>
...
WHERE {
 ?cultureResults
   ddo:hasCulture ?culturing;
   ddo:hasIdentified_Bacteria ?bug;
   ddo:hasDrug ?drug;
   ddo:hasAntibioticTestedResult ?antibiogramResult.
 ?culturing
   ddo:hasSample_Type ?sampleTypeCode;
   ddo:hasResult_Date ?resultDate.
 ?bug skos:notation ?uniProtCode.
 ?drug skos:notation ?atcCode.
FILTER (?uniProtCode =
  "1280"^^<http://eulersharp.sourceforge.net/2003/03swap/bioSKOSSchemes#
  uniProtTaxonomyDT>)
  # Staphylococcus aureus
FILTER (?atcCode =
  "J01XA01"^^<http://www.agfa.com/w3c/2009/clinicalSKOSSchemes#
  atc20090101DT>)
  # vancomycin_
```

```
FILTER (?sampleTypeCode =
  "119295008"^^<http://www.agfa.com/w3c/2009/clinicalSKOSSchemes#
  sct20080731DT>
  # Specimen obtained by aspiration
  ||...
  "258531008"^^<http://www.agfa.com/w3c/2009/clinicalSKOSSchemes#
  sct20080731DT>)
FILTER ("2007-01-01T00:00:00"^^<http://www.w3.org/2001/XMLSchema#dateTime>
<= ?resultDate && ?resultDate < "2008-01-01T00:00:00"^^
<http://www.w3.org/2001/XMLSchema#dateTime>)}
```

Lors de l'exécution de la requête, le résultat obtenu est reçu par le vCDR de l'IP :

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix ddo: <http://debugit.spim.jussieu.fr/ddo#>.
@prefix biosko: <http://eulersharp.sourceforge.net/2003/03swap/
  bioSKOSSchemes#>.
@prefix clisko: <http://eulersharp.sourceforge.net/2003/03swap/
  clinicalSKOSSchemes#>.
# 1 instance of result set
<https://debugit.spim.jussieu.fr/resource/culture_result/1970194>
  a ddo:CultureResults;
  ddo:hasCulture <https://debugit.spim.jussieu.fr/resource/
  culture_normalized/56313>;
  ddo:hasIdentified_Bacteria <https://debugit.spim.jussieu.fr/
  resource/bacteria/129>;
  ddo:hasDrug <https://debugit.spim.jussieu.fr/resource/
  drug/2491>;
  ddo:hasAntibioticTestedResult ddo:Sensitive.
<https://debugit.spim.jussieu.fr/resource/culture_normalized/56313>
  ddo:hasResult_Date "2007-01-14T23:00:00"^^xsd:dateTime;
  ddo:hasSample_Type "119311002"^^clisko:sct20080731DT. # Catheter specimen
<https://debugit.spim.jussieu.fr/resource/bacteria/129>
  skos:notation "1280"^^biosko:uniProtTaxonomyDT.
<https://debugit.spim.jussieu.fr/resource/drug/2491>
```

```
  skos:notation "J01XA01"^^clisko:atc20090101DT.
```

IP utilisera les règles définies dans le Knowledge Repository pour demander à EULER de réécrire les résultats en DCO. Voici un jeu de règles de réécriture :

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.
@prefix bisko: <http://eulersharp.sourceforge.net/2003/03swap/
bioSKOSSchemes#>.
@prefix clisko: <http://eulersharp.sourceforge.net/2003/03swap/
clinicalSKOSSchemes#>.
@prefix ddo: <https://debugit.spim.jussieu.fr/ddo#>.

# ANTIBIOGRAM
{ ?antibiogram a ddo:CultureResults;
  ddo:hasAntibioticTestedResult ?result}
=>
{ ?antibiogram biotop:encodes ?susceptibility.
  ?susceptibility
    a dco:MicrobiologicalSusceptibility;
    biotop:qualityLocated ?result}.

# BACTERIAL ANTIBIOGRAM ANALYSIS
{ ?susceptibilityTest
    a dco:AntimicrobialSusceptibilityTest;
    biotop:hasOutcome ?antibiogram.
  ?antibiogram biotop:encodes ?susceptibility.
  ?susceptibility a dco:MicrobiologicalSusceptibility}
=>
{ ?analysis
    a dco:BacterialAntibiogramAnalysis;
    biotop:hasParticipant ?antibiogram;
    biotop:hasOutcome [biotop:encodes ?susceptibility]}.
```

Et voici le jeu de données précédent réécrit après l'exécution des règles par EULER :

```
@prefix biotop:<http://purl.org/biotop/biotop.owl#>.
@prefix dco: <http://purl.org/imbi/dco/dco#>.

# 1 instance of converted result set
_:sk78_1_1
  a dco:BacterialAntibiogramAnalysis;
  biotop:hasParticipant <https://debugit.spim.jussieu.fr/resource/
  culture_result/1970194>;
  biotop:hasOutcome _:sk79_1_1.
<https://debugit.spim.jussieu.fr/resource/culture_result/1970194>
  a dco:Antibiogram;
  biotop:spatiallyRelatedTo _:t78_1_1;
  biotop:encodes [a dco:Vancomycin].
_:sk79_1_1 biotop:encodes _:sk39_1_1.
_:t78_1_1
  a biotop:TaxonQuality;
  biotop:qualityLocated
    [a dco:SpeciesStaphylococcusAureusValueRegion].
_:sk39_1_1
  a dco:MicrobiologicalSusceptibility;
  biotop:qualityLocated [a dco:Sensitive].
```

Une fois que les données ont été intégrées sémantiquement, des règles d'analyse vont permettre de calculer les éléments de la requête calculables, par exemple les pourcentages. Nous ne détaillerons pas cette partie dans ce document car il sort du cadre d'intégration sémantique de données réparties. Il restera intéressant de savoir que d'autres règles sont utilisées afin de calculer les pourcentages à partir de plusieurs jeux de données. Enfin, les résultats sont visualisés sur une interface homme-machine.

7.4.2.3 Résultats

Les résultats de l'intégration des données sont à ce jour, en train d'être évalués dans divers hôpitaux européens. Nous avons cependant mis en place sur les données historiques des interfaces web afin d'exploiter les données. Une interface web permettant à l'utilisateur de définir ses propres gadgets a été mise en oeuvre, et il est possible d'observer les données de résistance dans les sites européens et de

les confronter, ce qui n'aurait qu'un intérêt limité sachant que les résistances s'acquièrent assez localement. La figure 7.9 montre une courbe de résistance sur l'année 2007 pour le site HEGP d'E.Coli aux Fluroquinolones. Nous pouvons observer une certaine corrélation entre la courbe HEGP et la courbe agrégée.

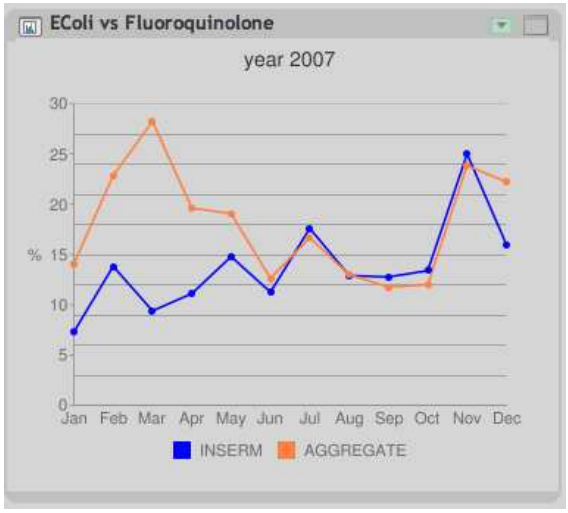


FIGURE 7.9 – Gadget montrant un graphique de résistance de E.Coli aux Fluoroquinolones au endpoint de l'INSERM sur l'année 2007 ainsi que les données agrégées des autres sites.

Ces résultats seront exploités à l'HEGP par les microbiologistes afin de créer un rapport annuel d'évolution de la résistance en interne (le rapport est aujourd'hui constitué manuellement). Ce rapport met en avant une analyse détaillée de l'évolution de résistances au cours de chaque année, et doit servir à aider les médecins à prescrire différemment. Idéalement, l'outil DebugIT devrait être interfacé directement dans le DPI de l'hôpital ce qui permettrait aux médecins de connaître les taux de résistance au moment de la prescription et d'adapter celles-ci. A ce stade de validation du projet, cette option n'est pas prévue à l'HEGP.

7.4.3 Validation de l'approche de médiation sémantique

Bien que la validation des outils et de l'approche DebugIT soit en cours, nous avons été capable de valider notre approche en comparant les résultats de résistance aux antibiotiques sur les années 2001-2007 rétrospectivement. Nous n'avons pu avoir accès aux données brutes de l'HEGP (nombre de cas) mais seulement aux pourcentages de résistance. Nous avons donc comparé les pourcentages de résistance au cours de cette période pour une sélection de couples antibiotique-bactérie. Nous indiquons par ailleurs le support sur lequel ces résultats sont obtenus (nombre de

	2001	2002	2003	2004	2005	2006	2007
HEGP (rapport)	74%	70%	71%	70%	71%	70%	69%
DebugIT (SPARQL endpoint)	73%	72%	72%	71%	72%	70%	69%
Support	1805	2714	2895	2782	2837	2860	2725

TABLE 7.5 – Taux de sensibilité de E.Coli à la Trimethoprime à l'HEGP sur une période de 6 ans.

tests sur l'année).

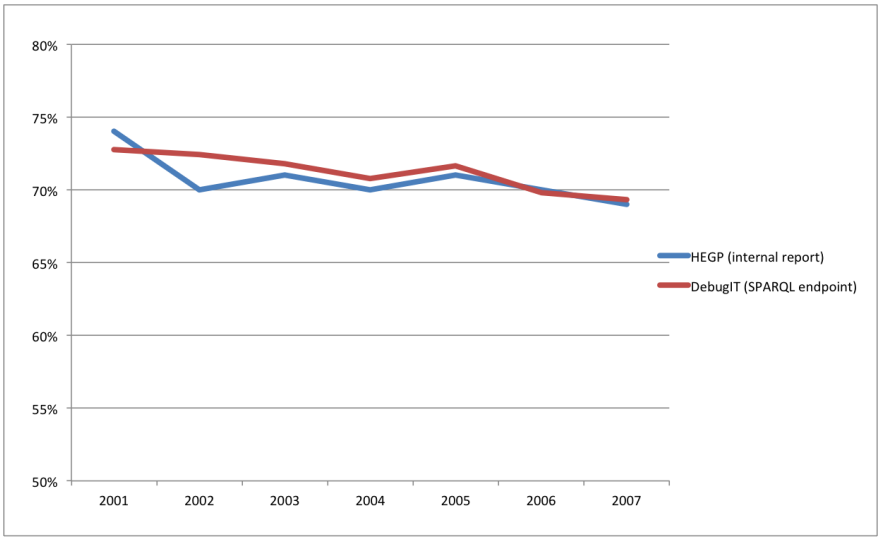


FIGURE 7.10 – Graphique de taux de sensibilité de E.Coli à la Trimethoprime à l'HEGP sur une période de 6 ans.

Comme nous pouvons le constater les résultats bien qu'étant du même ordre, diffèrent légèrement. Plusieurs pistes sont évoquées à ce jour pour expliquer ces différences. Tout d'abord, les données extraites du rapport microbiologiste sont extraites directement du système qui génèrent les résultats. Ensuite, des algorithmes experts dédoublonnent les cas de doublons possibles. Un doublon dans cette étude est une même souche sur un même patient qui serait re-testé dans la même année. Enfin, notre expérimentation (qualité, ajout de sémantique et médiation) peut avoir généré des erreurs. Nous remarquons cependant que les pourcentages sont généralement plus élevés. L'erreur de doublon semble donc la plus cohérente. Nous pouvons voir par contre que dans quelques cas, nous avons un taux de résistance plus faible. Cela peut être dû aux doublons aussi (dans le cas où il y a plus de cas de résistance) pour la même souche.

	2001	2002	2003	2004	2005	2006	2007
HEGP (internal report)	73%	85%	90%	91%	91%	91%	89%
DebugIT (SPARQL endpoint)	73%	86%	89%	91%	92%	91%	91%
total	1244	2404	2853	2780	2838	2850	2727

TABLE 7.6 – Taux de sensibilité de E.Coli à la Cefixime à l'HEGP sur une période de 6 ans.

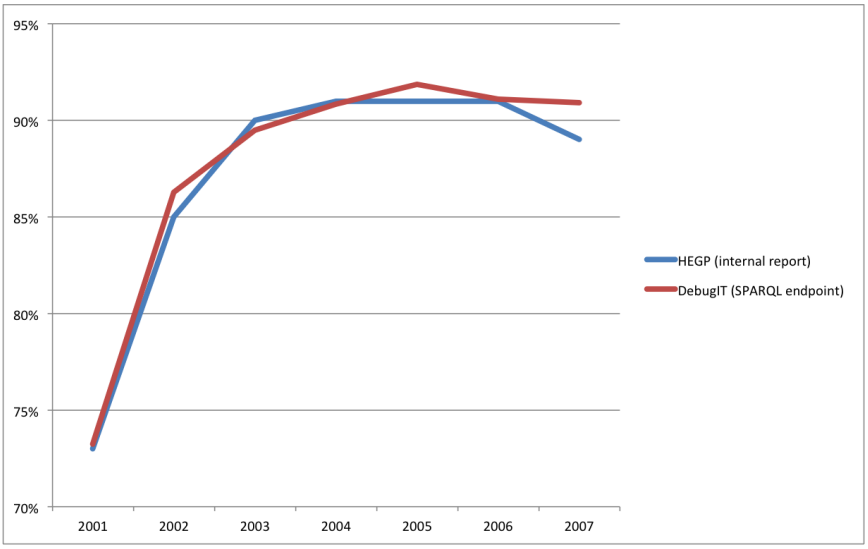


FIGURE 7.11 – Graphique de taux de sensibilité de E.Coli à la Cefixime à l'HEGP sur une période de 6 ans.

7.5 Conclusion

Nous avons présenté dans ce chapitre les expérimentations effectuées dans le cadre de l'interopérabilité de données d'analyse microbiologiques en Europe. Localement, nous avons proposé une méthode d'évaluation de la qualité des données de l'HEGP pour l'interopérabilité dans le contexte défini par le projet, et nous avons évalué cette source. Nos résultats ont été validés durant l'exploitation des données, lorsque nous avons dû partager les données de l'HEGP avec le reste du projet.

	2001	2002	2003	2004	2005	2006	2007
HEGP (internal report)	77%	72%	78%	76%	81%	81%	81%
DebugIT (SPARQL endpoint)	77%	72%	78%	78%	80%	82%	82%
total	728	1092	1125	1803	2726	2662	2730

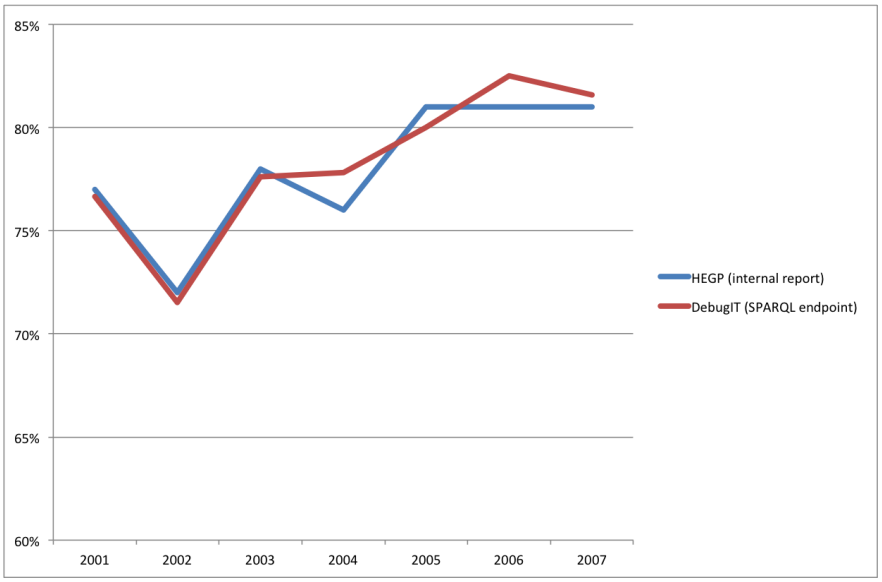


FIGURE 7.12 – Graphique de taux de sensibilité de E.Coli à la Chloramphenicol à l'HEGP sur une période de 6 ans.

Ensuite, nous avons proposé une modélisation dimensionnelle basée sur HL7 afin d'offrir une vue d'analyse de nos données qui soit plus facilement interopérable (comparativement au modèle de données non HL7).
Nous avons par la suite mis en oeuvre une ontologie de données (DDO) que nous avons ensuite enrichie à l'aide d'informations relatives à la qualité de l'information, et aux vocabulaires utilisés en base. La DDO est mise en oeuvre dans tous les CDR européens du projet.
Puis, nous avons défini les spécifications de la plateforme d'interopérabilité de DebugIT, tout d'abord fonctionnelles et ensuite techniques. Avec la collaboration d'Agfa Healthcare, nous avons développé une plateforme d'interopérabilité sémantique basée sur des règles de réécriture de requêtes et d'agrégation de données. Cette plateforme a la particularité de permettre une large montée en charge sur le temps. Nous avons validé les résultats de notre intégration en confrontant nos données aux données des experts locaux. L'évaluation de l'usage de la plateforme en routine est actuellement en cours et finira en Juillet 2012.

Troisième partie

Conclusion Générale

Chapitre 8

Conclusions, discussions et perspectives

"L'homme sage n'est pas comme un vase ou un instrument qui n'a qu'un usage ; il est apte à tout." - Confucius.

Sommaire

8.1 Introduction 178
8.2 Partage et modèles 178
8.3 Partage et qualité 180
8.4 Partage et sémantique 180
8.5 La plateforme d'interopérabilité sémantique 182
8.6 Contributions Personnelles 182
8.7 Conclusion générale 183

Les technologies de l'information ont évoluées. La récente démonstration d'IBM en la matière est surprenante[1]*, l'ajout de connaissances (ontologies) dans la base de faits de Watson (leur machine capable de battre des champions de Jeopardy, un jeu télévisé) est très intéressante et montre que la connaissance du monde, lorsqu'elle est correctement modélisée permet d'apporter à la machine des capacités de réponse dans un jeu de questions-réponses. Dans le domaine biomédical, la connaissance du monde n'est pas encore figée. Il reste complexe de représenter l'amplitude de cette connaissance difficile à modéliser et donc, à partager de manière consensuelle. Nous sommes donc dans un problème complexe, où le domaine est flou et où la machine a du mal à trouver son usage. Watson semble pouvoir gérer un monde fermé, nous pensons que la connaissance biomédicale évolue dans un monde ouvert, où l'inconnu est un concept.*

1. http://www.lemondeinformatique.fr/actualites/lire-le-supercomputer-watson-d-ibm-defie-les-candidat
html

Le titre de cette thèse est large. Trop, diront certains. Je l'ai choisi volontairement. Car achever l'interopérabilité des systèmes d'information est, en fin de compte, résoudre le problème de la machine pensante. En effet, pour que deux machines se comprennent, il faut qu'elles aient la capacité d'interpréter des messages. Or, comme le disait Albert Einstein, "La connaissance s'acquiert par l'expérience, tout le reste n'est que de l'information.", les machines gèrent habituellement de l'information, sans connaissance.

8.1 Introduction

Nous avons, tout au long de ce manuscrit de thèse, élaboré sur un sujet : le partage d'informations biomédicales. Pour partager de l'information, il faut savoir où la trouver, comment y accéder et interpréter son contenu. Nous nous sommes rapidement rendus compte que la connaissance liée à la qualité de l'information à partager était nécessaire pour avancer dans ce domaine. Nous avons ensuite travaillé avec les standards (modèles, vocabulaires, ontologies) aidant à favoriser l'interopérabilité. Il est cependant naïf de penser que si nous parlions tous le même langage, nous serions capables de nous comprendre totalement et sans ambiguité. C'est pourquoi nous avons ensuite abordé le problème par la mise en oeuvre d'une plateforme étant capable de se servir de la sémantique pour avancer dans la problématique générale. Chemin faisant, nous avons rencontré des verrous, posé des hypothèses, pour en fin de compte apporter une vision générale sur la problématique d'interopérabilité dans le domaine particulier de l'information biomédicale. Nous allons aborder la discussion et la conclusion de nos travaux par le biais de la problématique de l'interopérabilité suivant les axes des représentations, des langages et des connaissances.

8.2 Partage et modèles

Pouvoir échanger des données ou des connaissances est possible lorsqu'on sait où et comment aller chercher ces données dans un premier temps. Puis, vient rapidement la question de la façon d'extraire les données dans la structure dans laquelle elle est stockée. Que ce soit une base de données ou bien un message, la problématique reste identique, il faut savoir qui est où, et comment extraire l'information. Nous avons abordé dans cette thèse cette problématique sous différents angles. Tout d'abord, nous avons éclairé le lecteur concernant les structures de

stockage existantes pour les données ou des connaissances. Historiquement, ces structures ont suivi deux axes d'évolution, le modèle relationnel de Codd, et le modèle de type graphe pour la représentation sémantique. Nous avons ensuite vu que l'Internet, de part sa structure même, commence à mettre sur le devant de la scène les structures de type graphe. Certains diront que ce n'est pas important, que le modèle de stockage pourrait même se simplifier à l'extrême et que l'important c'est le contenu. Il n'en reste pas moins que ces structures sont les garantes de la performance d'un système et de sa capacité à être interrogé, donc interopérable. En effet, plus un modèle de données stocke les données de manière implicite (l'approche EAV par exemple), plus la connaissance humaine sera nécessaire pour construire des requêtes et plus celles-ci seront difficiles à mettre en oeuvre. De l'autre côté, plus la structure de stockage sera explicite, plus lourde seront les mises à jour de ces structures.

La communauté du web nous a récemment apporté des représentations de structure permettant le stockage de données semi-structurées (XML) ainsi qu'un modèle simple de stockage permettant de représenter d'abord de la connaissance, puis des données : RDF. L'apparence simplicité du modèle RDF pourrait permettre une adoption plus large, même dans le monde des bases de données. Beaucoup d'avancées ont été faites au cours des 20 dernières années dans le monde des bases de données. D'un point de vue de la robustesse et de la performance. RDF et les graphes ne sont pas encore reconnus au niveau industriel et pour des raisons valables de performance. Beaucoup d'efforts sont cependant mis en oeuvre pour tenter de pallier cette problématique de performance actuellement [2].

Dans le domaine de l'information médicale, il est important de se rendre compte de la difficulté de créer et de maintenir des modèles d'informations standards. L'initiative HL7 v3 démontre la complexité du domaine, et, alors que HL7 v3 se veut être un standard permettant l'interopérabilité des données qu'il véhicule, le consortium ne parvient pas à l'imposer et les implémentations réelles d'HL7 v3 restent rares. De plus, la structuration même du RIM d'HL7 v3 semble poser problème pour achever une réelle interopérabilité, et ce même après une récente proposition de création d'un ensemble de services pour l'interopérabilité (Service Aware Interoperability Framework)[Landgrebe 2011]. De l'autre côté, des projets comme i2b2 connaissent un franc succès, grâce notamment, à un modèle d'information performant et simple à appréhender. Il n'en reste pas moins que le modèle générique de l'approche n'est pas aisément interrogeable et qu'il ne peut représenter

2. http://lod2.eu/BlogPost/528-call-for-participation-bsbm-3-1-benchmarks-on-the-lisa-cluster.html

les causalités entre les événements médicaux d'un même patient (pour les raisons évoquées précédemment).
Nous avons contribué en essayant de prendre le meilleur des deux visions. Un modèle issu des standards HL7 v3 que nous avons adapté afin qu'il garde son expressivité, sa performance et sa facilité d'interrogation. Notre approche n'est cependant pas facilement évolutive. Mais elle permet de gagner en interopérabilité en permettant l'interfaçage plus aisé via des connecteurs HL7 v3.

8.3 Partage et qualité

La notion de qualité dans un contexte de partage de l'information est essentielle pour les usages que nous ferons de l'information échangée. Dans le domaine médical surtout, les systèmes d'information décisionnels doivent pouvoir donner aux personnels médicaux des informations fiables. Nous avons remarqué que cela était difficile. En effet, autant qu'il est difficile de construire des terminologies justes et complètes, il est très complexe pour un professionnel de santé d'entrer des données sans erreurs dans le système d'information. Nous pensons donc qu'il est nécessaire, tout comme nous l'avons fait pour décrire la sémantique des sources d'information, d'exprimer la qualité de l'information que nous partageons. L'information toujours plus importante et accessible, doit, avant d'être partagée, qualifiée. Nous avons proposé une grille de lecture de ce qu'est la qualité de l'interopérabilité d'une source de données. Et nous avons appliqué cette grille dans un cadre local tout d'abord. Nous avons ensuite proposé une intégration de cette grille dans un cadre plus général, mais une réelle implémentation n'a pas encore été effectuée.

8.4 Partage et sémantique

L'homme utilise différentes natures de connaissances pour échanger de l'information. La connaissance liée au décodage des mots, des images ou des sons. La connaissance liée à l'environnement, à l'histoire de la personne. Et d'autres. Ces réseaux de connaissances, que l'on appelle d'ailleurs réseaux de part la typologie des neurones, sont en permanence mis à contribution pour faire évoluer d'autres réseaux. Des milliards d'opérations sont effectuées en même temps, à l'inverse de la machine qui est capable d'effectuer des milliers de milliards d'opérations à la chaîne. On comprend alors que la machine doit évoluer pour pouvoir faire face et traiter la complexité des réseaux sémantiques. Nous ne parlons pas d'une évolution de puis-

sance, mais plutôt de la manière dont elle doit traiter l'information.
Nous savons qu'il existe plusieurs natures de connaissances nécessaires au traitement de l'information. Des connaissances sur la localisation des données, sur la structure de stockage, sur les vocabulaires utilisés, sur le langage d'interrogation. Puis des connaissances sur le domaine, que l'on organisera suivant des concepts pères fondateurs (top-ontologie) ou non. Une fois ces connaissances formalisées, il faut les relier. Les utiliser dans des usages particuliers. Ces usages sont généralement définis. Dans le cadre de cette thèse, et de l'expérimentation effectuée au cours du projet DebugIT, nous avons pris le partit de formaliser au mieux les connaissances des données, de leur localisation, et du domaine. Nous avons ensuite créé des applicatifs capables d'utiliser cette connaissance pour des usages définis par l'homme. Nous avons pu parfois utiliser la connaissance formelle pour aider à l'interopérabilité, par exemple pour déduire des requêtes lors du processus d'intégration de données, ou bien pour l'adaptation des résultats obtenus à chaque source vers un formalisme commun. Nous avons pu aussi utiliser la connaissance pour effectuer des regroupements de données (subsomption). Mais beaucoup reste à faire.
La sémantique est une étape nécessaire pour aller vers un partage d'information mais ne suffit pas. Tout d'abord car il y a autant de sémantiques qu'il y a d'observateurs ou d'usages. Les langages de représentation actuels (ontologies) bien qu'extrêmement puissants, ne semblent pas être suffisants pour être capables de tout modéliser, pour tous les usages. Il va donc falloir inventer des systèmes capables d'utiliser plusieurs modélisations d'un même monde. Cela ne veut pas dire qu'il faut arrêter l'effort de convergence, car il améliore souvent la qualité du domaine modélisé. Mais il n'existera certainement jamais d'ontologie unique. Là encore, la machine doit aider l'homme, non le contraindre. C'est dans ce sens que nous avons développé la plateforme d'interopérabilité.
Les technologies sémantiques proposées sont en nombre grandissant. Et l'interopérabilité entre elles n'est pas toujours possible[Gacia 2011]. OWL,OWL2, OBO, SWRL, N3, etc. ont autant de formalismes, de langages et de représentations différentes de la connaissance. Dans un monde interopérable, les systèmes d'information sémantiques doivent pouvoir être interopérables entre eux. En améliorant l'interopérabilité sémantique, on ne doit pas perdre l'interopérabilité technique et syntaxique de ces mêmes systèmes.

8.5 La plateforme d'interopérabilité sémantique

L'interoperabilité nécessite de l'outillage. La plateforme informatique que nous avons imaginé, spécifié, et qui a été ensuite mis en oeuvre par Agfa Healthcare, répond aux exigences d'interopérabilité suivantes :

- Elle utilise des représentations de connaissances pour opérer (ontologie de données, ontologie décrivant les vocabulaires, vocabulaires, ontologies de domaine, ontologies utilisateur, etc.). Elle est capable d'opérer avec n'importe quelle ontologie de domaine et de données.
- Elle n'impose pas de vue globale.
- Elle permet de relier des vues différentes d'une même réalité grâce à des règles.
- Elle peut gérer n sources de données.
- Elle peut gérer n requêtes pré-définies.
- Elle ne peut pas encore s'auto-gérer pour répliquer la connaissance des liens entre les sites, elle est pour le moment centralisée.

Nous savons déjà que cette plateforme va évoluer à travers de nouveaux projets européens, et nous souhaitons qu'elle puisse être partagée dans la communauté car elle adresse une problématique universelle. La plateforme est aujourd'hui opérationnelle et complètement paramétrable par des ressources formelles. En travaillant sur les interfaces de paramétrage, elle pourrait intégrer des données dans divers domaines, et à grande échelle.

8.6 Contributions Personnelles

Les contributions de ce travail de thèse sont plurielles. Tout d'abord, nous avons proposé un cadre de lecture de la qualité de l'information biomédicale pour l'interopérabilite : le triangle de la qualité de l'information. Nous avons ensuite mis en oeuvre et testé cette méthode en l'intégrant au niveau de l'expression du modèle d'une source de données. Nous avons défini les spécifications de la plateforrme d'interopérabilité de DebugIT qui a été mis en oeuvre par la société Agfa Healthcare. Nous avons proposé une formalisation d'une ontologie de données qui donne lieu en ce moment à une implémentation d'automatisation de sa construction. Nous avons proposé, pour réduire le problème de réécriture de requêtes, d'intégrer une notion de vocabulaire dans l'ontologie de données. Nous avons ainsi réduit le problème d'utilisation d'une ontologie de domaine pour la médiation de données. Nous avons aussi proposé une généralisation du problème de réécriture de requêtes dans le cadre de l'utilisation d'une ontologie de domaine comme médiateur. Nous n'avons pas pu valider ce travail. Nous avons aussi été force de proposition pour la constitution d'un

modèle dimensionnel standardisé qui est actuellement implémenté et validé. Nous avons par ailleurs proposé une architecture d'entrepôt de données dimensionnels sémantique couplant les avantages des SGBD relationnels et de la formalisation riche d'une ontologie respectant certaines propriétés.

D'un point de vue plus fondamental, nous avons mis en avant tout au long de cette thèse des questionnements. L'interopérabilité sémantique d'abord. Nous avons montré que ce terme était souvent mal employé dans beaucoup de communautés, l'informatique médicale comprise. Dire que HL7 (par exemple) permet l'interopérabilité sémantique revient à dire que l'anglais permet de se comprendre. Il en va de même pour toute ressource "standardisante". La capacité qu'ont deux systèmes de se comprendre ne réside pas tant dans le mapping entre deux concepts que dans leur capacité à interpréter. Les ontologies sont bien évidemment des référentiels d'interprétation nécessaires, mais elles ne suffisent pas, il faut pouvoir les traiter, et utiliser pleinement les capacités de raisonnement qu'elles offrent dans ce contexte d'interopérabilité. Nous avons aussi soulevé le problème de la relation non matérialisée de manière persistante et interrogeable dans les modèles de données courants. Nous sommes convaincus que la relation doit être matérialisée afin de réduire l'implicite que renferment aujourd'hui les bases de données (ce qui favorise la création de cimetières de données). Enfin, nous avons montré qu'il était possible d'avoir plusieurs visions du monde, plusieurs usages de ses visions, et pour autant mettre en oeuvre une plateforme informatique sachant réconcilier et gérer ses visions, sans en imposer une.

8.7 Conclusion générale

L'internet nous pousse à réétudier la manière dont nous concevons l'information et comment nous y avons accès. En effet, le caractère hautement hétérogène du réseau où tout le monde peut contribuer, pousse la communauté du web sémantique à trouver des méthodes et des formalismes pour tenter d'interconnecter cette masse de données. L'utilisation des outils et méthodes issus de cette communauté dans le domaine de la santé est, pour nous, adaptée. Evidemment, le but premier de cette communauté est de faire communiquer des systèmes aux usages différents, sur un même réseau. Nous pourrions alors penser que ces méthodes et outils ne sont pas adaptés, car le monde médical est plus fermé. Mais comme nous l'avons vu, l'information médicale est aussi très hétérogène, tant par sa nature, que par sa complexité de modélisation. Or, le monde de l'information scientifique ou médicale n'évolue pas en silos. Il faut trouver des moyens d'interconnecter les données. Aussi,

penser le maillage de l'information biomédicale comme un tout est pour nous la seule manière d'avancer et d'espérer un jour atteindre un niveau acceptable d'interopérabilité en santé. Pour construire ce maillage, il faut que chaque maillon exprime au mieux l'information qu'il partage. En utilisant des langages et des formalismes d'expression explicites. Ce n'est qu'à partir de là que nous pourrons imaginer des systèmes "intelligents" pour traiter cette information. Nous sommes donc convaincus que l'approche ascendante est une méthode de construction réaliste et essentielle pour achever l'interopérabilité.

Le domaine de la santé est aujourd'hui vu comme un domaine qui doit s'informatiser à grande échelle. Plusieurs raisons nous poussent vers cela. Si on met de côté l'aspect purement d'économie d'échelle (qui n'est pas prouvé) il semble aujourd'hui nécessaire d'aider la profession de santé à avoir un accès aux données des patients rapidement (même lorsqu'il vient de l'extérieur), à avoir un meilleur accès à la connaissance (nous vivons dans une aire de sur-spécialisation de la médecine) de manière à leur libérer du temps pour mieux soigner et prendre de bonnes décisions. La complexité de l'information médicale est un frein évident. Car les systèmes d'information "classiques" s'accommodent mal de celle-ci. Il nous apparaît clair aujourd'hui que les systèmes d'information de santé doivent évoluer pour intégrer la sémantique, non comme une évolution add-hoc, mais comme un changement fondamental d'architecture au coeur du système d'information.

La question des standards reste centrale dans notre travail. Comme nous l'avons vu, ces standards aident à mieux qualifier les concepts médicaux. Ils sont cependant toujours spécialisés. Et rendent compte d'une vision particulière du monde pour un usage définit. Ces standards évoluent aussi, et doivent s'interconnecter. Ils proposent une solution séduisante au problème de l'interopérabilité, mais nous pensons qu'il sera toujours difficile de s'accorder sur des standards. Aussi, nous croyons qu'il est nécessaire et important d'offrir l'outillage nécessaire à affronter la "jungle" des standards. Tout autant que la "jungle" des cimetières de données d'ailleurs. Nous ne pourrons forcer l'utilisation de standards, ni garantir de la qualité de ceux-ci. Nous pensons qu'il faut être pragmatique et faciliter l'interopérabilité déjà au niveau des données. Pour la question des standards, comme de la qualité, nous pensons qu'il faudra intégrer la notion d'incertitude et d'intervalle de confiance, même au niveau de l'interconnexion des données. Nous pensons donc qu'il faut partir des données, et imaginer des systèmes d'interconnexion semi-automatiques de type alignement. Il faut aujourd'hui travailler plus sur la notion d'alignement sémantique (aidé ou non par des ontologies de domaine). L'expression du modèle de données en RDF est essentiel car il permet de travailler conjointement avec les données et les méta-

données. La plateforme d'interopérabilité pourra permettre la mise en oeuvre de ces alignements.

Les projets européens de recherche en eSanté offrent un terrain d'échange privilégié en Europe. Les équipes de recherche en informatique biomédicale cherchent activement des solutions innovantes afin d'interconnecter les systèmes de santé pour encore mieux connecter les hommes et leurs connaissances. Le domaine de la santé est un monde ouvert où l'information se doit de circuler entre professionnels, mais aussi entre professionnels et patients. Fluidifier les échanges électroniques permettra, nous l'espérons, de donner aux professionnels de santé l'information dont ils ont besoin, quand ils en ont besoin, pour mieux nous soigner.

"Les machines un jour pourront résoudre tous les problèmes, mais jamais aucune d'entre elles ne pourra en poser un !" - Albert Einstein.

Bibliographie

[Abrial 1974] JR Abrial. *Data Semantics*. IFIP Working Conference of Data Base Management, J.W. Klimbie and K.L. Koffeman edition, North Holland Publishing Company, pages 1–60, 1974. 25

[Antonioletti 2005] M Antonioletti, M Atkinson, R Baxter et A Borley. *The design and implementation of Grid database services in OGSA-DAI*. Concurrency and Computation : Practice & Experience, Jan 2005. 83

[Arts 2002] DGT Arts, NF De Keizer et GJ Scheffer. *Defining and improving data quality in medical registries : a literature review, case study, and generic framework*. Journal of the American Medical Informatics Association, vol. 9, no. 6, page 600, 2002. 45

[Baneyx 2007] Audrey Baneyx. Construire une ontologie de la pneumologie, 2007. 4, 5, 35

[Baril 2003] X Baril et Z Bellahsène. *Designing and Managing an XML Warehouse*. XML Data Management : Native XML and XML-Enabled Database Systems, 1st Edition, pages 455–474, 2003. 23

[Basili 2002] Victor R. Basili, Gianluigi Caldiera et H. Dieter Rombach. *Encyclopedia of Software Engineering, Experience Factory*. 2002. 45

[Beale 2002] T Beale. *Archetypes : Constraint-based domain models for future-proof information systems*. OOPSLA 2002 workshop on behavioural semantics, 2002. 105

[Beneventano 2001] Domenico Beneventano, Sonia Bergamaschi, Francesco Guerra et Maurizio Vincini. *The MOMIS approach to Information Integration*. 2001. 83

[Berners-Lee 2001] T Berners-Lee. *Scientific publishing on the semantic web*. Nature, 2001. 33, 59

[Berners-Lee 2009] T Berners-Lee. *Linked Data-The Story So Far*. International Journal on Semantic Web and . . . , 2009. 51

[Berti-Equille 2005] L Berti-Equille et F Moussouni. *Quality-aware integration and warehousing of genomic data*. Proceedings of the 10th international conference on . . . , Jan 2005. 44

[Beuscart 2011] Régis Beuscart. *PSIP : an overview of the results and clinical implications*. Studies in health technology and informatics, vol. 166, pages 3–12, Jan 2011. 98

[Bizer 2007] C Bizer et R Cyganiak. *D2RQ-Lessons Learned.* Proceedings of the W3C Workshop on . . . , Jan 2007. 122

[Bourigault 1999] D Bourigault et M Slodzian. *Pour une terminologie textuelle.* Terminologies Nouvelles, 1999. 108

[Bourigault 2004] D Bourigault et N Aussenac-Gilles. *Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas.* Revue d'Intelligence . . . , 2004. 35

[Boussaid 2006] O Boussaid, R Messaoud, R Choquet et S Anthoard. *X-Warehousing : An XML-Based Approach for Warehousing Complex Data.* Proceedings of the 10 thEast-European Conference on Advances Databases, 2006. 21, 23

[Bracchi 1976] G Bracchi, P Paolini et G Pelagatti. *Binary Logical Associations in Data Modelling.* IFIP Working Conference on Modelling in data Base Management Systems, ed. G.M. Nijssen, North-Holland, pages 125–148, 1976. 25

[Broekstra 2002] J Broekstra, A Kampman et F Van Harmelen. *Sesame : A generic architecture for storing and querying rdf and rdf schema.* The Semantic Web—ISWC . . . , Jan 2002. 122

[Brown 2000] PJB Brown et P Sonksen. *Evaluation of the quality of information retrieval of clinical findings from a computerized patient database using a semantic terminological model.* Journal of the American Medical Informatics Association, vol. 7, no. 4, pages 392–403, 2000. 109

[Charlet 2002] J Charlet. *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales.* 2002. 35, 37

[Choquet 2009] R Choquet, D Kalra et MC Jaulent. *Specification of an Inter-Operability Platform for the integration and exploitation of distributed clinical data.* American Medical Informatics Association Annual Conference, page 1–1, Mar 2009. 129

[Choquet 2011] Remy Choquet, Daniel Karlsson, Daniel Schober, Philip Daumke, Patrick Ruch, Douglas Teodoro, Giovanni Mels, Ariane Assele, Dirk Colaert, Christian Lovis, Marie-Christine Jaulent, Hans Cools et Jos De Roo. *Exchanging biomedical information at large scale using the Semantic Web.* In XXIII International Conference of the European Federation for Medical Informatics (MIE), 2011. 159

[Codd 1970] E Codd. *A relational model of data for large shared data banks.* Communications of the ACM, Jan 1970. 18

[Codd 1979] E Codd. *Extending the database relational model to capture more meaning.* ACM Transactions on Database Systems (TODS), Jan 1979. 18

[Coloma 2011] P Coloma, M Schuemie et G Trifiro. *Combining electronic healthcare databases in Europe to allow for largescale drug safety monitoring : the EUADR Project.* Pharmacoepidemiology and Drug Safety, Jan 2011. 98

[Davidson 1996] S. B. Davidson, C. Overton, V. Tannen et L. Wong. *BioKleisli : A Digital Library for Biomedical Researchers.* 1996. 83

[De Roo 2002] J De Roo. Euler proof mechanism. Internet : http ://www. agfa. com/w3c/euler, 2002. 40

[de Saussure 1916] F de Saussure. *Writings in general linguistics.* books.google.com, Jan 1916. 38

[Deen 1987] S M Deen, R R Amin et M C Taylor. *Data Integration in Distributed Databases.* IEEE Transactions on Software Engineering, vol. SE-13, no. 7, pages 860–864, Juillet 1987. 56

[Degoulet 1998] Patrice Degoulet et Marius Fieschi. Informatique médicale. Elsevier Masson, 1998. 4

[Deming 2000] W Deming. *Out of the Crisis.* MIT Press, Cambridge, Jan 2000. 44

[Dinu 2007] Valentin Dinu et Prakash Nadkarni. *Guidelines for the effective use of entity-attribute-value modeling for biomedical databases.* International Journal of Medical Informatics, vol. 76, no. 11-12, pages 769–779, Octobre 2007. 24

[Dittrich 1986] K Dittrich. *Object-oriented database systems (extended abstract) : the notions and the issues.* OODBS : international workshop on Object-oriented database systems, pages 2–4, Jan 1986. 20

[Fellegi 1969] I Fellegi et A Sunter. *A theory for record linkage.* Journal of the American Statistical Association, Jan 1969. 43

[Gacia 2011] R Gacia. *Perspectives in Semantic Interoperability.* International Workshop on Semantic Interoperability IWSI 2011, 2011. 181

[Garcia-Molina 1997] Hector Garcia-Molina, Yannis Papakonstantinou, Dallan Quass, Anand Rajaraman, Yehoshua Sagiv, Jeffrey Ullman, Vasilis Vassalos et Jennifer Widom. *The TSIMMIS Approach to Mediation : Data Models and Languages.* Journal of Intelligent Information Systems, vol. 8, pages 117–132, 1997. 83

[Goasdoué 1999] Francois Goasdoué, Marie-Christine Rousset et et al. *The Use of CARIN Language And Algorithms For Information Integration : The PIC-*

SEL System. In INTELLIGENT INFORMATION INTEGRATION WORKSHOP ASSOCIATED WITH ECAI'98 CONFERENCE, 1999. 83

[Goldberg 2008] SI Goldberg, A Niemierko et A Turchin. *Analysis of Data Errors in Clinical Research Databases*. AMIA Annual Symposium Proceedings, vol. 2008, page 242, 2008. 45

[Gomez-Perez 2004] A Gomez-Perez et M Fernández-López. Ontological engineering, 2004. 41

[Gorman 2006] MM Gorman et A Tolk. *Next generation data interoperability : it's all about metadata*. ... of Fall Simulation Interoperability ..., 2006. 52, 80

[Grémy 1987] François Grémy. Informatique médicale, 1987. 3

[Gruber 1993] TR Gruber. *A translation approach to portable ontology specifications*. Knowledge acquisition, 1993. 37

[Guarino 1997] N Guarino. *Some organizing principles for a unified top-level ontology*. ... (AAAI-97) in the Spring Symposium on Ontological ..., 1997. 42

[Guarino 1998] N Guarino. *Formal ontology in information systems*. 1998. 41

[Hanke 1999] Jens Hanke, Gerrit Lehmann, Peer Bork et Jens G Reich. *Associative Database of protein sequences*. Bioinformatics, vol. 15, no. 9, pages 741–748, Sep 1999. 26

[Horrocks 2003] I Horrocks et PF Patel-Schneider. *From SHIQ and RDF to OWL : The making of a web ontology language*. Web semantics : science, 2003. 40

[Hümmer 2003] W Hümmer, A Bauer et G Harde. *XCube : XML for data warehouses*. ... 6th ACM international workshop on Data warehousing and OLAP (DOLAP), pages 33–40, Jan 2003. 23

[Ikeda 1997] M Ikeda et R Mizoguchi. *A causal time ontology for qualitative reasoning*. ... JOINT CONFERENCE ON ..., 1997. 42

[Inmon 1995] W Inmon. What is a data warehouse ?, volume 1. Prism Tech Topic, Jan 1995. 21

[Kerr 2007] K Kerr, A Norris et R Stockdale. *Data Quality Information and Decision Making : A Healthcare Case Study*. Proc. 18th Australasian Conference on Information Systems, Jan 2007. 45

[Kimball 1995] R Kimball et K Strehlo. *Why decision support fails and how to fix it*. ACM SIGMOD Record, Jan 1995. 21

[Kimball 2002] R Kimball et Margy Ross. The data warehouse toolkit. John Wiley and Sons, Jan 2002. 21

[Krogstie 2010] J Krogstie, O Lindland et G Sindre. *Towards a deeper understanding of quality in requirements engineering*. Advanced Information Systems Engineering, 2010. 44

[Landgrebe 2011] J Landgrebe et B Smith. *The HL7 Approach to Semantic Interoperability*. International Conference in Biomedical Ontology, 2011. 179

[Lee 2006] Thomas Lee, Yannick Pouliot, Valerie Wagner, Priyanka Gupta, David Stringer-Calvert, Jessica Tenenbaum et Peter Karp. *BioWarehouse : a bioinformatics database warehouse toolkit*. BMC Bioinformatics, vol. 7, no. 1, page 170, 2006. 82

[Leibniz 1886] GW Leibniz. Nouveaux essais sur l'entendement humain (avant-propos et livre premier)., 1886. 63

[Lovis 2006] Christian Lovis, Stéphane Spahni, Nicolas Cassoni-Schoellhammer et Antoine Geissbuhler. *Comprehensive management of the access to a component-based healthcare information system*. Studies in health technology and informatics, vol. 124, pages 251–256, 2006. 6

[Missier 2008] Paolo Missier. *MODELLING AND COMPUTING THE QUALITY OF INFORMATION IN E-SCIENCE*. PhD Thesis, pages 1–227, Janvier 2008. 126

[Mizoguchi 1995] R Mizoguchi et J Vanwelkenhuysen. *Task ontology for reuse of problem solving knowledge*. Towards Very Large ..., 1995. 41

[Moody 2003a] DL Moody. *Measuring the quality of data models : an empirical evaluation of the use of quality metrics in practice*. Proceedings of the Eleventh European Conference on Information Systems, ECIS, 2003. 44

[Moody 2003b] DL Moody et GG Shanks. *Improving the quality of data models : empirical validation of a quality management framework*. Information Systems, vol. 28, no. 6, pages 619–650, 2003. 44, 109

[Nassis 2004] V Nassis, R Rajugan, T Dillon et W Rahayu. *Conceptual Design of XML Document Warehouses*. DaWaK, pages 1–14, Jan 2004. 23

[Naumann 2000] F Naumann et C Rolker. *Assessment methods for information quality criteria*. Proceedings of the International Conference on ..., Jan 2000. 44

[O'Connor 2010] Martin J O'Connor et Amar Das. *Semantic reasoning with XML-based biomedical information models*. Studies in health technology and informatics, vol. 160, no. Pt 2, pages 986–990, 2010. 123

[Ouagne 2010] D Ouagne, N Nadah et D Schober. *Ensuring HL7-based information model requirements within an ontology framework.* Studies in health ..., 2010. 111, 114, 118

[Peralta 2008] V Peralta. *Data Quality Evaluation in Data Integration Systems.* hal.archives-ouvertes.fr, Jan 2008. 44

[Pierra 2005] G Pierra, H Dehainsala, Y Ameur et L Bellatreche. *Base de données à base ontologique : principe et mise en oeuvre.* Ingénierie des Systèmes d'Information (ISI), Jan 2005. 82

[Pillai 1987] Sushil V. Pillai, Ramanatham Gudipati et Leszek Lilien. *Design issues and an architecture for a heterogenous multidatabase system.* In Proceedings of the 15th annual conference on Computer Science, CSC '87, pages 74–79, New York, NY, USA, 1987. ACM. 83

[Pokorny 2001] J Pokorny. *Modelling stars using XML.* Proceedings of the 4th ACM international workshop on Data and OLAP, Jan 2001. 23

[Redman 1996] T Redman. *Data quality for the information age.* slac.stanford.edu, Jan 1996. 43

[Reynaud 2002] C Reynaud et G Giraldo. *Vers l'automatisation de la construction de systèmes de médiation pour le commerce électronique.* Journées de l'Action Spécifique Web, 2002. 58

[Roussopoulos 1975] N Roussopoulos et J Mylopoulos. *Using semantic networks for data base management.* Proceedings of the 1st International Conference on Very Large Databases, pages 144–172, Jan 1975. 32

[Ruelland 2003] Alan Ruelland, Marie-Christine Jaulent, Mario Ota, Bruno Frandji et Patrice Degoulet. *Pragmatic objects modeling environment for Electronic Health Records Systems.* Studies in health technology and informatics, vol. 95, pages 328–333, 2003. 25

[Schober 2010] Daniel Schober, Martin Boeker, Hans Cools, Douglas Teodoro, Jessica Bullenkamp, Nadia Nadah, R Choquet et Stefan Schulz. *The DebugIT Core Ontology : semantic integration of antibiotics resistance patterns.* 13th International Congress on Medical Informatics, 2010. 121, 130

[Schulz 2006] S Schulz, E Beisswanger et U Hahn. *From GENIA to BIOTOPTowards a Top-Level Ontology for Biology.* ... on Formal Ontology ..., 2006. 42

[Schulz 2007] Stefan Schulz, Boontawee Suntisrivaraporn et Franz Baader. *SNOMED CT's problem list : ontologists' and logicians' therapy suggestions.* Stu-

dies in health technology and informatics, vol. 129, no. Pt 1, pages 802–806, 2007. 105

[Schulz 2010] Stefan Schulz, Daniel Schober, Christel Daniel et Marie-Christine Jaulent. *Bridging the semantics gap between terminologies, ontologies, and information models.* Studies in health technology and informatics, vol. 160, no. Pt 2, pages 1000–1004, 2010. 116

[Sears 2005] Cynthia L Sears. *A dynamic partnership : celebrating our gut flora.* Anaerobe, vol. 11, no. 5, pages 247–251, Octobre 2005. 91

[Shah 2005] Sohrab Shah, Yong Huang, Tao Xu, Macaire Yuen, John Ling et BF Francis Ouellette. *Atlas - a data warehouse for integrative bioinformatics.* BMC Bioinformatics, vol. 6, no. 1, page 34, 2005. 82

[Sheth 1990] A Sheth et J Larson. *Federated database systems for managing distributed, heterogeneous, and autonomous databases.* ACM Computing Surveys (CSUR), Jan 1990. 57

[Shironoshita 2008] EP Shironoshita, YR Jean-Mary, RM Bradley et MR Kabuka. *semCDI : a query formulation for semantic data integration in caBIG.* Journal of the American Medical Informatics Association, vol. 15, no. 4, page 559, 2008. 84

[Smith 2006] Barry Smith et Werner Ceusters. *HL7 RIM : an incoherent standard.* Studies in health technology and informatics, vol. 124, pages 133–138, 2006. 74

[Spackman 2004] KA Spackman. *Examining SNOMED from the perspective of formal ontological principles : Some preliminary analysis and observations.* Whistler, 2004. 105

[Stevens 2000] R. D. Stevens, P. G. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. Paton, C. A. Goble et A. Brass. *TAMBIS : transparent access to multiple bioinformatics information sources.* Bioinformatics, vol. 16, pages 184–186, 2000. 83

[Stolba 2006] N Stolba, M Banek et A Tjoa. *The Security Issue of Federated Data Warehouses in the Area of Evidence-Based Medicine.* ARES, Jan 2006. 83

[Stonebraker 1996] M Stonebraker et D Moore. Object relationnal dbms, the next great ware. Morgan and Faufman Publishers, ISBN 1-55860-397-2, 1996. 21

[Stroetmann 2009] VN Stroetmann, D Kalra, P Lewalle et JM Rodrigues. Semantic interoperability for better health and safer healthcare. ... Commission (available at ..., 2009. 80

[Taylor 2006] P Taylor. *Patient-centred records*. 2006. 75

[Thalhammer 2001] T Thalhammer et M Schrefl. *Active data warehouses : complementing OLAP with analysis rules*. Data & Knowledge Engineering, 2001. 56

[Töpel 2008] T Töpel, B Kormeier et A Klassen. *BioDWH : A data warehouse kit for life science data integration*. Journal of Integrative . . . , Jan 2008. 82

[Traver 2011] V Traver et R Faubel. *Personal Health : The New Paradigm to Make Sustainable the Health Care System*. Biomedical Engineering Systems and Technologies : Third International Joint Conference, BIOSTEC 2010, Jan 2011. 98

[Uschold 1995] M Uschold. *Towards a methodology for building ontologies*. Workshop on basic ontological issues in knowledge . . . , 1995. 38

[Van Heijst 1997] G Van Heijst et AT Schreiber. *Using explicit ontologies in KBS development*. International Journal of Human . . . , 1997. 41

[Vassiliadis 1999] Panos Vassiliadis, Mokrane Bouzeghoub et Christoph Quix. *Towards Quality-Oriented Data Warehouse Usage and Evolution*. vol. 1626, pages 164–179, 1999. 45

[Vicente 2006] Miguel Vicente, John Hodgson, Orietta Massidda, Tone Tonjum, Birgitta Henriques-Normark et Eliora Z Ron. *The fallacies of hope : will we discover new antibiotics to combat pathogenic bacteria in time ?* FEMS microbiology reviews, vol. 30, no. 6, pages 841–852, 2006. 96

[Wand 1996] Yair Wand et Richard Y. Wang. *Anchoring data quality dimensions in ontological foundations*. Commun. ACM, vol. 39, pages 86–95, November 1996. 44

[Wang 1998] R Wang. *A product perspective on total data quality management*. Communications of the ACM, Jan 1998. 43, 44, 108

[Weikum 1999] G Weikum. *Towards guaranteed quality and dependability of information systems*. Proc. of the Conf. Datenbanksysteme inB*uro, Technik und Wissenschaft, 1999. 44

[Williams 2001] S Williams. *The associative model of data*. Journal of Database Marketing, Jan 2001. 25

[Wisniewski 2003] MF Wisniewski, P Kieszkowski, BM Zagorski, WE Trick, M Sommers et RA Weinstein. *Development of a clinical data warehouse for hospital infection control*. Journal of the American Medical Informatics Association, vol. 10, no. 5, pages 454–462, 2003. 43

[Zweigenbaum 1999] P Zweigenbaum. *Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances.* Innovation Stratégique en . . . , 1999. 35

Liste des publications :

Choquet R, Daniel C, Boussaid O, Jaulent MC. Etude méthodologique comparative de solutions d'entreposage de données de santé à des fins décisionnelles. 9ème Conférence internationale sur la science des systèmes de santé. 2008.

Teodoro D, Choquet R, Pasche E, Gobeill J, Daniel C, Ruch P, Lovis C. Biomedical Data Management : A Proposal Framework. Studies in health technology and informatics. Proceedings of Medical Informatics European conference. 2009.

Teodoro D, Pasche E, Wipfli R, Gobeill J, Choquet R, Daniel C, Ruch P, Lovis C. 2009. Integration of biomedical data using federated databases. Proceedings of the 22th conference of swiss medical informatics. 2009.

Choquet R, Kalra D, Jaulent MC. Specification of an Inter-Operability Platform for the integration and exploitation of distributed clinical data. American Medical Informatics Association Annual Conference. 2009.

Schober D, Boeker M, Bullenkamp J, Huszka C, Depraetere K, Teodoro D, Nadah N, Choquet R, Daniel C, Schulz S. The DebugIT core ontology : semantic integration of antibiotics resistance patterns.. Studies in health technology and informatics. 160(Pt 2) :1060-4. 2010.

Ouagne D, Nadah N, Schober D, Choquet R, Teodoro D, Colaert D, Schulz S, Jaulent M-C, Daniel C. Ensuring HL7-based information model requirements within an ontology framework.. Studies in health technology and informatics. 160(Pt 2) :912-6. 2010.

Choquet R, Qouiyd S, Ouagne D, Pasche E, Daniel C, Boussaïd O, Jaulent M-C. The Information Quality Triangle : a methodology to assess clinical information quality.. Studies in health technology and informatics. 160(Pt 1) :699-703. 2010.

Choquet R, Teodoro D, Mels G, Assele A, Pasche E, Ruch P, Lovis C, Jaulent

M-C. Partage de données biomédicales sur le web sémantique. Ingénierie de la Connaissance (IC2010). 2010.

Choquet R, Qouiyd S, Pasche E, Daniel C, Boussaïd O, Jaulent M-C. Un modèle de connaissances pour mesurer la qualité d'une source d'information. 21e Journées francophones d'Ingénierie des Connaissances (IC2010). 2010.

Assele A, Mels G, Choquet R, Jaulent M-C. Utilisation d'une ontologie comme source de données pour répondre à des questions cliniques. 21e Journées francophones d'Ingénierie des Connaissances (IC2010). 2010.

Daniel C, Choquet R, Assele A, Enders F, Daumke P, Jaulent M-C. Comparing the DebugIT dashboards to national surveillance systems. International Conference on Prevention and Infection Control (ICPIC 2011). Poster. 2011.

Teodoro D, Choquet R, Schober D, Mels G, Pasche E, Ruch P, Lovis C. Interoperability driven integration of biomedical data sources. XXIII International Conference of the European Federation for Medical Informatics (MIE). 2011.

Choquet R, Daniel C, Grohs P, Douali N, Jaulent M-C. 2011. Monitoring the emergence of antibiotic resistance using the technology ot the DebugIT platform in the HEGP context. ICPIC 2011. Poster. 2011.

Résumé : Le volume de données disponibles dans les systèmes d'information est de plus en plus important et pour autant, nous n'avons jamais autant essayer d'interconnecter cette information pour en extraire de la connaissance sans véritable succès généralisable. L'origine du problème est multiple. Tout d'abord, l'information est représentée dans des structures différentes. Ensuite, les vocabulaires utilisés pour exprimer les données sont hétérogènes. Enfin, la qualité de l'information est souvent trop mauvaise pour utiliser une information et en déduire des connaissances. Ce diagnostic est d'autant plus vrai dans le cadre du partage d'information dans le domaine biomédical où il reste difficile de s'entendre sur des représentations (structures et vocabulaires) pivots d'un domaine de la médecine, et donc où il apparaît difficile de résoudre le problème du partage d'information par l'imposition de standard de codage et de structuration de l'information. Plus récemment, l'introduction de la sémantique dans des processus de partage d'information, nous offre la possibilité de mettre en oeuvre des représentations pivots indépendantes de la structuration ou du nommage d'une donnée.

Cette thèse s'inscrit dans cette problématique de partage de données biomédicales dans le cadre de l'évaluation de l'évolution de la résistance des bactéries aux antibiotiques en Europe. L'hypothèse générale de travail que nous proposons est la suivante : comment partager de l'information biomédicale de manière non ambigüe, en temps réel, et à la demande en Europe. Cette hypothèse pose diverses problématiques que nous abordons dans ce mémoire. La problématique de la qualité des données. Celle de la représentation des données à travers leur structure, leur vocabulaire et de leur sémantique. Nous aborderons aussi les problèmes d'alignement de données aux ontologies de domaine et de la fédération de données aidée d'ontologie. Enfin, nous présenterons un système d'interopérabilité sémantique basé sur des règles qui aborde le problème d'alignement sémantique de systèmes hétérogènes appliqué à notre domaine. Nous discuterons finalement de l'apport de la sémantique pour le partage d'information et des limites des outils et méthodes actuels.

Mots clés : Interopérabilité sémantique, entrepôts de données, informatique médicale, médiation sémantique de données, ontologies, qualité de données, intégration de données, modèles d'information, web sémantique, standards, réécriture de requête, règles, logique de description, raisonnement, projet européen, DebugIT.

Abstract : The amount of available data in information systems is constantly increasing and more and more efforts have been made in trying to interconnect this data in order to gain knowledge or meaning. Yet, these attempts at interconnecting such data have never been satisfactory enough when it comes to using the information at a wider scale. The origins of such difficulties are manifold. First, information is represented in many different structures. Second, the vocabulary used to express data is heterogeneous. Finally, the quality of the information is often too poor to be used and to withdraw any knowledge from it. Such observation applies specifically to the biomedical area where it is still difficult to agree on a common and shared representation (structures and vocabulary) concerning a particular sub-domain of the medical field. It would appear difficult in such a context to solve the problem of information sharing by imposing standard coding and standard information models. More recently, the introduction of semantics in the process of information sharing enables us to setup pivots representations which are independent from the structure or from the naming of the data.

This thesis deals with the problematics of biomedical information sharing in the study of antibiotics resistance evolution to bacteries in Europe. Our general working hypothesis is : how can we share biomedical information in Europe in a non ambiguous way, in a fast way, and on demand ? Many issues are raised by this working hypothesis. We will deal with the issue of the quality of the data, the issue of the representation of data through their structure, their vocabulary, their semantics. We will also address the problems of alignment of data with domain ontologies. And the problem of data mediation helped with domain ontologies. We will then present a system of semantic interoperability based on rules which addresses the problem of semantic alignment of heterogeneous systems applied to our domain. Finally we will discuss how semantics can contribute to the improvement of information sharing and we will also discuss the limits of the current tools and methods.

Keywords : Data Integration, semantic interoperability, data warehouse, medical informatics, semantic mediation, ontologies, data quality, information models, semantic web, standards, query rewriting, rules, description logic, reasoning, european project, DebugIT.