



# Apport de la modélisation ontologique pour le partage des connaissances en psychiatrie

Marion Richard

## ► To cite this version:

Marion Richard. Apport de la modélisation ontologique pour le partage des connaissances en psychiatrie. Complexité [cs.CC]. Université Pierre et Marie Curie - Paris VI, 2017. Français. NNT : 2017PA066202 . tel-01798089

**HAL Id: tel-01798089**

**<https://theses.hal.science/tel-01798089>**

Submitted on 23 May 2018

**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.

**THÈSE DE DOCTORAT DE  
L'UNIVERSITÉ PIERRE ET MARIE CURIE**

SPÉCIALITÉ : INFORMATIQUE MÉDICALE

ÉCOLE DOCTORALE 393  
PIERRE LOUIS DE SANTÉ PUBLIQUE À PARIS :  
ÉPIDÉMIOLOGIE ET SCIENCES DE L'INFORMATION BIOMÉDICALE

*Présentée par*

**Marion RICHARD**

Pour obtenir le grade de

**DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE**

**Apports de la modélisation ontologique pour le partage des  
connaissances en psychiatrie**

Soutenue le 19 septembre 2017 devant le jury composé de :

Directeur	Jean CHARLET	- AP-HP, INSERM/LIMICS
Co-encadrant	Xavier AIMÉ	- INSERM/LIMICS
Rapporteurs	Giuseppe BERIO	- Univ. Bretagne Sud, IRISA
	Michel WALTER	- CHU de Brest, Univ. Bretagne Occidentale
Examineurs	Jean-Gabriel GANASCIA	- Univ. Pierre et Marie Curie, LIP6
	Marie-Odile KREBS	- CH Sainte-Anne, Univ. Paris Descartes
	Marion ROBIN	- Institut Mutualiste Montsouris
	Romain SICOT	- Hôpital Fernand Widal



## REMERCIEMENTS

Mes premiers remerciements s'adressent à Messieurs Jean Charlet et Xavier Aimé, qui ont dirigé et soutenu ce travail de recherche. Leur expérience dans le domaine de l'informatique médicale, leurs conseils, leur patience et leur implication ont favorisé la réussite de cette thèse.

Je tiens également à remercier l'équipe du service hospitalo-universitaire (SHU) du centre hospitalier Sainte-Anne. En particulier Marie-Odile Krebs avec qui le projet a pris forme. Mais également les psychiatres et psychologues cliniciens qui ont participé à la validation de l'ontologie : Boris Chaumette, Macarena Cuenca, Olivier Gay, Célia Mam Lam Fook, Yannick Morvan et Dominique Willard. Cette collaboration a participé à la valeur de ce travail de recherche.

Je remercie également Sylvie Sulzman pour Termina, Cyril Grouin pour Medina et Antonio Lossio pour BioTex. Je les remercie pour leur temps et leur aide dans la manipulation de leurs outils.

Je remercie Giuseppe Berio et Michel Walter d'avoir accepté de faire profiter ma thèse de leur expérience en informatique et en psychiatrie. La vision de leur domaine respectif m'a permis d'affiner le travail présenté dans ce manuscrit. Je remercie également Jean-Gabriel Ganascia, Marion Robin et Romain Sicot d'avoir accepté de faire partie de mon jury de thèse.

Mes remerciements s'adressent également à l'équipe du LIMICS, Marie-Christine Jaulent pour son accueil dans l'unité et les chercheurs des diverses antennes du laboratoire. J'ai une pensée particulière pour Troskah Farnia, Mary Shi, Sonia Cardoso, Felipe Melo, Lamine Traore, Adrien Ugon pour les conversations internationalement riches et les éclats de rire. Je remercie Adriane Podasca et Isabelle Verdier pour leur aide administrative. Je remercie tous les chercheurs du laboratoire avec qui j'ai échangé des conversations passionnantes et qui ont toujours su me donner des conseils pertinents dans les moments de doute et de tension. Mon passage au LIMICS a été une réussite grâce à toutes les personnes que j'ai pu rencontrer au cours de ces quatre années.

Pour conclure, un grand merci à ma famille et mes amis qui ont su me soutenir et s'intéresser à cet immense projet.



# Table des matières

<b>Introduction</b>	<b>13</b>
Contexte de recherche et problématiques . . . . .	13
Enjeux et objectifs . . . . .	16
Contributions à la croisée de plusieurs disciplines . . . . .	17
Organisation du manuscrit . . . . .	19
 <b>I État de l'art</b>	 <b>21</b>
<b>1 L'organisation des connaissances</b>	<b>23</b>
1.1 Les systèmes d'organisation de la connaissance (SOC) . . . . .	24
1.2 L'approche de la sémiologie et de la sémiotique pour la modélisation des connaissances . . . . .	27
1.3 Un point sur les classifications utilisées en psychiatrie . . . . .	31
1.4 Synthèse . . . . .	38
 <b>2 Système de représentation ontologique</b>	 <b>41</b>
2.1 L'ontologie informatique . . . . .	42
2.2 Contexte : le Web Sémantique (WS) . . . . .	45
2.3 Les composants de la modélisation ontologique . . . . .	55
2.4 Synthèse . . . . .	57
 <b>3 Construction d'ontologies informatiques</b>	 <b>59</b>
3.1 Engagements méthodologiques des ontologues et recommandations pour la construction d'ontologies . . . . .	60
3.2 Développement du modèle par approche ascendante (bottom-up) . . . . .	62
3.3 Développement du modèle par approche descendante (top-down) . . . . .	73
3.4 Développement du modèle par approche hybride (bottom-up et top-down) . . . . .	77
3.5 Discussions sur les méthodes de construction d'ontologies . . . . .	78
3.6 La modularité ontologique : l'ergonomie au service du développement du modèle. . . . .	79
3.7 Synthèse . . . . .	83

<b>4</b>	<b>L'art de valider une ontologie</b>	<b>85</b>
4.1	La définition des critères de validation . . . . .	86
4.2	Validation de la structure . . . . .	90
4.3	Validation de la sémantique . . . . .	93
4.4	Synthèse . . . . .	94
<b>II</b>	<b>Contributions scientifiques théoriques et pratiques</b>	<b>97</b>
<b>5</b>	<b>Construction du module ontologique « facteurs sociaux et environnementaux des maladies psychiatriques » (OntoPsychiaFSE)</b>	<b>99</b>
5.1	Choix de ce module . . . . .	100
5.2	Présentation du corpus . . . . .	101
5.3	Méthode de construction par approche hybride . . . . .	104
5.4	Résultats . . . . .	108
5.5	Synthèse . . . . .	113
<b>6</b>	<b>Construction du module ontologique « maladies psychiatriques »</b>	<b>115</b>
6.1	Choix du module . . . . .	116
6.2	Méthode de construction des deux modules . . . . .	117
6.3	Résultats . . . . .	121
6.4	Les limites du module d'alignement des classifications . . . . .	129
6.5	Synthèse . . . . .	130
<b>7</b>	<b>La validation de l'ontologie sur les facteurs sociaux et environnementaux avec la méthode interactive LOVMI</b>	<b>133</b>
7.1	Validation de la structure de l'ontologie sur les facteurs sociaux et environnementaux . . . . .	134
7.2	Validation sémantique de l'ontologie sur les facteurs sociaux et environnementaux . . . . .	139
7.3	Proposition de la méthode LOVMI pour la validation d'ontologies . . . . .	145
7.4	Synthèse . . . . .	148
	<b>Conclusion</b>	<b>149</b>
	<b>Bibliographie</b>	<b>157</b>
<b>III</b>	<b>Annexes</b>	<b>173</b>
<b>A</b>	<b>Cadre administratif de la thèse</b>	<b>175</b>
<b>B</b>	<b>Illustration des théories de la sémiotique de Hébert et Dumont-Morin [2012]</b>	<b>177</b>
<b>C</b>	<b>Illustration de la dixième version de la Classification statistique Internationale des Maladies et des problèmes de santé connexes</b>	<b>179</b>
<b>D</b>	<b>Les « Constructs/Subconstructs » des RDoC</b>	<b>181</b>

<b>E</b>	<b>Nuages des liens entre les données ouvertes</b>	<b>183</b>
<b>F</b>	<b>Extrait du code source de la page de la BnF répertoriant les « signets » en format Dublin core</b>	<b>187</b>
<b>G</b>	<b>Les critères de qualité définis par Gherasim <i>et al.</i> [2012]</b>	<b>189</b>
<b>H</b>	<b>Anonymisation des comptes rendus d'hospitalisation</b>	<b>191</b>
	H.1 Le respect de la confidentialité pour l'exploitation de données médicales . .	191
	H.2 Anonymisation du corpus avec l'aide du logiciel MEDINA . . . . .	195
	H.3 Synthèse . . . . .	198
<b>I</b>	<b>Table de correspondance entre les étiquettes de TREE TAGGER et de MELT</b>	<b>199</b>
<b>J</b>	<b>Extrait du fichier xml résultat de l'extraction avec YATEA</b>	<b>201</b>
<b>K</b>	<b>Extrait du fichier xml résultat de l'extraction avec BIOTEX</b>	<b>203</b>
<b>L</b>	<b>Arborescence conceptuelle de l'ontologie sur les facteurs sociaux et environnementaux</b>	<b>205</b>
<b>M</b>	<b>Arborescence conceptuelle construite à partir des termes extraits des comptes rendus d'hospitalisation</b>	<b>209</b>





# Liste des figures

1	Interdisciplinarité de nos contributions. . . . .	18
1.1	Le signe linguistique selon Ferdinand de Saussure. . . . .	28
1.2	Le signe linguistique selon Charles Sanders Peirce. . . . .	28
2.1	The Semantic Web - Not a piece of cake...by Benjamin Nowack . . . . .	47
2.2	Un graphe RDF représentant un triplet : deux nœuds (Sujet et Objet) reliés entre eux par une relation (le Prédicat) Smith <i>et al.</i> [2004]. . . . .	47
3.1	Des données vers un modèle structuré Aussenac-Gilles et Charlet [2010]. . .	63
3.2	Méthodologie de King et Uschold [1995] d'après Fernández-López et Gómez- Pérez [2002]. . . . .	64
3.3	Methontology . . . . .	66
3.4	La méthode ARCHONTE [Bachimont <i>et al.</i> , 2002]. . . . .	67
3.5	De modèles génériques vers un modèle adapté [Aussenac-Gilles et Charlet, 2010]. . . . .	74
3.6	Le modèle de connaissances de COMMONKADS d'après Schreiber [2000]. . .	75
3.7	La méthodologie de SENSUS illustrée dans Swartout <i>et al.</i> [1996]. . . . .	76
5.1	Hiérarchie des relations modélisées dans l'ontologie pour décrire des liens entre les concepts. . . . .	112
6.1	Exemple d'une extraction de termes candidats avec le logiciel BIOTEX. . . .	120
6.2	Diagramme des alignements entre CIM-10, DSM IV TR et DSM 5. . . . .	124
6.3	Hiérarchie conceptuelle issue de la catégorisation du DSM 5 dans la fenêtre de gauche. Exemple d'alignement de code avec la CIM 10 et de catégorie avec le DSM IV TR à droite. . . . .	125
7.1	Résultats de l'évaluation de notre module <i>facteurs sociaux et environnemen- taux des maladies psychiatriques</i> par OOPS !. . . . .	136
7.2	Résultat de la requête (sous Protégé) permettant d'extraire les concepts sans PrefLabel en anglais. . . . .	138
7.3	Schéma modélisant le cadre méthodologique LOVMI. . . . .	147

B.1	Les diverses appellations du signe linguistique, page 248 de Hébert et Dumont-Morin [2012]. . . . .	177
C.1	Cette image illustre la description d'une catégorie dans la CIM-10. Un code est associé à un libellé et à des indications thérapeutiques. Par exemple sur cette image, le code F00* est associé au libellé « Démence de la maladie d'Alzheimer ». Le code entre parenthèse G30.0-+ indique un lien avec le code diagnostique G30.0 qui est celui associé au libellé « Maladie d'Alzheimer à début précoce » . . . . .	179
C.2	Cette image illustre les différentes catégories regroupés dans le Chapitre 5 de la CIM-10 : « Troubles mentaux et du comportement ». Chacune de ces catégories principales est désigné par un intervalle de codes. . . . .	180
E.1	Les données liées en 2007. . . . .	183
E.2	Les données liées en 2009. . . . .	184
E.3	Les données liées en 2010. . . . .	184
E.4	Les données liées en 2017. . . . .	185
H.1	Association de deux clefs privées servant à crypter et décrypter un texte par [Quantin <i>et al.</i> , 2013]. . . . .	194
H.2	Procédure d'anonymisation mise en place dans le cadre de notre projet. . .	196
L.1	Extrait 2 de la hiérarchie conceptuelle de l'ontologie. . . . .	206
L.2	Extrait de la hiérarchie conceptuelle de l'ontologie. . . . .	207
L.3	Extrait de la hiérarchie conceptuelle de l'ontologie. . . . .	208
M.1	Premier niveau de hiérarchie conceptuelle de la branche « Acte médical » à gauche, et des branches « Diagnostic » et « Parcours de soins » à droite. . . .	209
M.2	Premier niveau de hiérarchie conceptuelle de la branche « Etat » à gauche, et de la branche « Trouble » à droite. . . . .	210
M.3	Premier niveau de hiérarchie conceptuelle de la branche « Observation de la condition du patient » à gauche, de la branche « Observation de l'évolution de la maladie » au centre, et de la branche « Observation médicale générale » à droite. . . . .	211

# Liste des tableaux

1.1	Présentation sommaire des différents SOC définis dans ce chapitre. . . . .	27
1.2	Analyse sémique de deux sous-types de la schizophrénie. . . . .	31
2.1	Présentation de la typologie adoptée dans ces travaux. . . . .	44
3.1	Calcul de la performance des ETC. . . . .	72
3.2	Résultats de l'analyse du corpus test par les ETC. . . . .	73
5.1	Comparaison des résultats des annotateurs morphosyntaxiques TREETAG-GER et MELT. . . . .	106
5.2	Métriques du module « Facteurs sociaux et environnementaux des maladies psychiatriques » . . . . .	110
5.3	Quelques statistiques sur le module « Facteurs sociaux et environnements des maladies psychiatriques » (OntoPsychiaFSE). Les nombres de classes incluent les classes à deux parents. . . . .	110
6.1	Répartition des codes selon leur fréquence d'apparition dans les CRH. . . .	121
6.2	Résultats de l'alignement selon le DSM 5. . . . .	122
6.3	Résultats de l'alignement avec OnaGUI. . . . .	122
6.5	Répartitions des alignements dans le modèle de connaissances des classifications. . . . .	123
6.4	Résultats de l'alignement transitif. . . . .	123
6.6	Alignement des codes et catégories relatifs à la « schizophrénie ». . . . .	126
6.7	Liste des 20 codes CIM-10 les plus fréquemment rencontrés dans les CRH de Sainte-Anne, alignés sur les catégories des DSM. . . . .	126
7.1	Résultats de l'analyse de OOPS ! sur le module <i>facteurs sociaux et environnementaux des maladies psychiatriques</i> . . . . .	137
7.2	Résultats de l'analyse de OOPS ! sur le module des <i>trouble psychiatriques</i> . .	137
7.3	ONTOPSYCHIA FSE avant la première phase de validation avec les acteurs. .	141
7.4	ONTOPSYCHIA FSE après la première phase de validation avec les acteurs. .	142
7.5	La taxonomie d'ONTOPSYCHIA FSE durant la deuxième phase de validation avec les acteurs. . . . .	143
7.6	Résultats de l'annotation des CRH avec ONTOPSYCHIA. . . . .	145

7.7	Présentation de la méthode Lovmi. . . . .	146
7.8	L'ontologie des « Facteurs sociaux et environnementaux des maladies psychiatriques » avant la validation avec les experts. . . . .	148
7.9	L'ontologie des « Facteurs sociaux et environnementaux des maladies psychiatriques » après la validation avec les experts. . . . .	148

# Introduction

## Contexte de recherche et problématiques

### Apport de la modélisation ontologique...

Le mot ontologie est emprunté au mot latin scientifique « *ontologia* »<sup>1</sup>. En philosophie, ce mot sert à désigner une branche fondamentale de la métaphysique, qui est selon Aristote « la science de l'être en tant qu'être ». L'ontologie s'occupe de ce qui existe, des propriétés générales de l'être. L'informatique a repris ce terme par analogie, pour nommer les artefacts qui permettent une représentation formelle de la connaissance, de ce qui existe dans le monde. Selon la définition consensuelle de Gruber [1995], une ontologie informatique est une formalisation d'une conceptualisation partagée. C'est un artefact qui permet la recherche sémantique, les raisonnements formels, ou encore l'intégration de données en vue d'une interopérabilité et d'une interprétation de ces données par un ordinateur.

Depuis une dizaine d'années, les méthodes de construction d'ontologies se sont fortement développées au travers du traitement automatique du langage (TAL) et de l'intérêt croissant pour les corpus de données volumineux. De nombreuses méthodes de construction semi-automatiques ont vu le jour telles que, ONTOLEARN [Velardi *et al.*, 2013], ARCHONTE [Bachimont *et al.*, 2002, Charlet *et al.*, 2006] ou TERMINAE [Aussenac-Gilles *et al.*, 2008]. Le développement de ces méthodes de construction d'ontologies, fondées sur l'extraction de termes spécialisés au sein de corpus a engendré un effacement progressif des acteurs du domaine et placé l'ontologue au centre du processus. Cependant, ces acteurs du domaine demeurent les détenteurs de la connaissance encyclopédique et pratique qui peut faire défaut à l'ontologue. On observe également que ces méthodes automatiques ont permis de développer des ontologies de taille plus importante, entraînant du même coup une plus grande difficulté à assurer une modélisation adéquate au domaine et aux formalismes ontologiques. La validation d'ontologies est par conséquent devenue une problématique à part entière de l'ingénierie des connaissances (IC). Dans un ouvrage dédié à cette problématique, Vrandečić [2009] a relevé trois scénarios qui justifient la validation et que nous résumons comme tels : (1) une ontologie adéquate permettra une meilleure réutilisation des données ; (2) les ontologues ont besoin de méthodes pour évaluer et valider leurs modèles, afin de les encourager à partager leurs résultats

---

1. Ce dernier est composé du grec *onto-*, tiré du grec ancien *ontos*, qui signifie « étant, ce qui est », et de *-logia*, tiré du grec ancien *logos* qui signifie « discours, traité »

avec la communauté et leur permettre de réutiliser avec confiance le travail des autres à leurs propres fins ; (3) les méthodes de validation d'ontologies permettent de vérifier automatiquement si les contraintes et les exigences sont remplies et de révéler ainsi les problèmes de plausibilité. Cela diminue les coûts d'entretien des ontologies. On note également dans la littérature, que la validation se définit sous deux aspects complémentaires : (1) la *validation structurelle* qui peut être réalisée automatiquement grâce au développement d'outils dédiés [Guarino et Welty, 2000, Fernández-López et Gómez-Pérez, 2002, Völker et al., 2008, Shearer et al., 2008, Schober et al., 2012, Poveda-Villalón et al., 2012], et (2) la *validation sémantique* qui peine encore à trouver des méthodes consensuelles [Pammer et al., 2010, Ghidini et al., 2012, Ressay-Bouidghaghen et al., 2013, Ben Abacha et al., 2013]. Notre recherche s'inscrit dans ces problématiques contemporaines liées au traitement et à l'utilisation de corpus volumineux pour la construction d'ontologies, ainsi que celles liées à la validation des ontologies.

### ...pour le partage des connaissances en psychiatrie

Le domaine médical est de plus en plus demandeur de systèmes fondés sur des ontologies, afin de permettre notamment le codage des actes médicaux ou l'aide à la décision. De nombreuses ontologies ont été réalisées à ce jour comme, par exemple ONTOLURGENCE pour la médecine d'urgence [Charlet et al., 2012, 2014] ou FOUNDATIONAL MODEL OF ANATOMY en anatomie [Rosse et al., 2003]<sup>2</sup>. Dans le domaine de la psychiatrie, de nombreux travaux se sont intéressés à la puissance de modélisation des ontologies, notamment dans les but de :

1. Gérer l'interopérabilité des données : En 2010, Kola et al. [2010] se sont intéressés aux problèmes d'hétérogénéité des données et au besoin d'interopérabilité dans le domaine de la psychiatrie. Les auteurs rappellent que les ontologies ont été utilisées avec succès pour résoudre les problématiques liées à l'interopérabilité des données dans d'autres champs de la biologie. Dans leur article, ils discutent les besoins de disposer d'une ontologie de domaine pour modéliser les psychoses. Ils proposent également une méthodologie pour créer une telle ontologie, mais d'un point de vue logique, sans intention de redéfinir des termes déjà existants.
2. Établir une terminologie consensuelle : Quelques années plus tard, Hastings et al. [2012] ont présenté une méthodologie pour développer une ontologie qui modélise les maladies mentales. Leur attention s'est portée sur l'adéquation de leur ontologie avec la Basic Formal Ontology (BFO<sup>3</sup>) qui est une top ontologie pour la recherche scientifique, et l'Ontology for General Medical Science (OGMS<sup>4</sup>) qui est une ontologie pour le domaine de la médecine générale. Les chercheurs ont analysé le DSM-5 et le modèle de Pies [2009]<sup>5</sup> pour définir les maladies mentales et identifier certaines confusions dans la terminologie psychiatrique. À partir de cette approche et de cette

2. On note également qu'un certain nombre de terminologies médicales sont présentées comme des ontologies formelles. Cela peut créer une confusion et des problèmes d'interprétation. Nous n'abordons pas cette difficulté dans nos travaux. Toutefois, en section 1.1.2, une présentation des principaux systèmes d'organisation de la connaissance permet de les différencier les uns par rapport aux autres.

3. <https://bioportal.bioontology.org/ontologies/BFO>

4. <https://bioportal.bioontology.org/ontologies/OGMS>

5. Le modèle de Pies définit cinq étapes pour décider de l'aspect pathologique d'un trouble, par exemple à partir de quel stade une addiction devient un trouble mental ?

méthode, ils ont développé deux ontologies : la Mental Functioning Ontology (MF) pour représenter les processus mentaux divisés en trois modules (cognition, perception, émotion), et la Mental Disease Ontology (MD) pour décrire et catégoriser les troubles mentaux. Ces réalisations sont le fruit de réflexions menées par **Ceusters et Smith [2010]** sur la définition d'une maladie mentale. Les chercheurs signalaient à cette occasion que le développement de leur ontologie visait à (1) servir de pont entre différentes classifications (à la manière de la Classification Internationale des Maladies - CIM) ; (2) permettre la collecte de données à partir de dossiers médicaux électroniques dont les modes de diagnostic diffèrent ; (3) résoudre certaines limites imposées par les classifications actuelles telles que les difficultés liées à la délimitation des symptômes ou à la comorbidité. Les objectifs de ce projet sont donc très proches de ceux visés par notre proposition d'alignement entre les classifications. Cependant, les méthodes de développement de nos modèles diffèrent. Nous ne basons pas nos analyses sur les mêmes classifications et pour ce qui nous concerne, nous ne visons pas à définir ce qu'est un trouble mental. Malgré de nombreuses recherches, nous n'avons pu trouver qu'un modèle très incomplet de leur ontologie<sup>6</sup>. Nous ne pouvons donc comparer nos travaux respectifs.

3. Permettre le raisonnement automatique : Plus récemment, en 2014, **Silva et al. [2014]** ont relancé la discussion autour de la représentation des diagnostics. L'objectif des travaux de ces chercheurs est quelque peu différent des travaux que nous venons de présenter. En effet, ils ne montrent pas d'intérêt particulier aux problèmes liés aux classifications en psychiatrie, mais s'intéressent plutôt aux difficultés liées à la modélisation de diagnostic. Ils ont ainsi développé un système basé sur une ontologie pour faciliter le diagnostic de troubles mentaux et permettre une description plus rationnel de ces diagnostics. Leur ontologie peut inférer des troubles selon les symptômes évoqués. Ils ont choisi le DSM-IV pour la modélisation. La partie *aide au diagnostique* du manuel est représentée sous forme de règles exprimées en OWL-DL et SWRL. La vérification de leur modèle par des psychiatres n'a pas donné de bons résultats. Les chercheurs souhaitent donc valider les règles en comparant les diagnostics inférés par leur système, à un ensemble de données patients.
4. Modéliser l'historique médical : La Family Health History Ontology (FHHO) **Peace et Brennan [2007]** n'a pas été conçue uniquement pour servir dans le domaine de la psychiatrie. Cette ontologie modélise différents liens familiaux, tels que les liens biologiques, d'adoption ou de reconstitution. Les ontologistes qui ont développé cette ontologie avaient pour but de proposer un cadre pour explorer (1) les interactions entre tous les membres de la famille et (2) les effets de ces interactions sur la santé et la maladie.

Environ une personne sur trois souffrira d'un trouble mental au cours de sa vie. Dans l'Union Européenne, 82,7 millions de personnes sont touchées chaque année. Malgré des classifications des troubles mentaux internationalement reconnues, la catégorisation des patients selon des critères diagnostiques reste problématique. En effet, les troubles sont dénotés par des syndromes qui possèdent des symptômes propres à une ou plusieurs catégories diagnostiques. Lorsqu'un patient présente des symptômes qui franchissent les

---

6. <https://raw.githubusercontent.com/jannahastings/mental-functioning-ontology/master/ontology/MFOMD.owl>



frontières des catégories descriptives et qui ne correspondent à aucune des sous-catégories diagnostiques, le recours au « NOS » *Not Otherwise Specified* permet d'attribuer une étiquette diagnostique et de ne pas laisser le patient exempt des codages. Par exemple, un patient présentant un ensemble de symptômes qui ne correspond pas exactement à une sous-catégorie diagnostique recevra effectivement un diagnostic général, dans le cadre des politiques de codage. Cependant, ce code ne sera pas représentatif en tout point de la réalité clinique<sup>7</sup> décrite dans le compte rendu d'hospitalisation (CRH). Ce type de recours au « NOS » marque une incohérence entre la description des troubles répertoriés dans les classifications, et la réalité clinique telle qu'elle est appréhendée par les professionnels de la santé mentale. En outre, l'analyse de la prévalence et de l'incidence des facteurs de risque sociaux et environnementaux des maladies est actuellement absente des classifications et manuels psychiatriques. Pourtant, cette analyse apparaît de plus en plus cruciale pour comprendre et traiter les troubles. Elle peut avoir des répercussions importantes sur les décisions thérapeutiques et politiques, sur la durée ou le coût de l'hospitalisation par exemple. Cette rupture entre les classifications et la réalité clinique souligne la nécessité d'améliorer : (1) les systèmes de modélisation et de description des troubles mentaux, ainsi que (2) notre capacité à détecter et comprendre l'impacte des facteurs de risque sociaux et environnementaux sur les troubles.

## Enjeux et objectifs

### La prise en compte des facteurs sociaux et environnementaux

La médecine de précision permet de définir des sous-groupes dans la population, à l'aide de marqueurs biologiques ou génétiques [Picard, 2014]. Cette stratification de la médecine est rendue possible grâce aux développements des nouvelles technologies, ainsi qu'aux progrès réalisés en génétique, neurosciences, cognition et santé mentale. Elle a pour but de soigner les patients en fonction des caractéristiques de leur pathologie et des spécificités génétiques et environnementales qui peuvent différer d'un individu à l'autre. Dès lors, le patient n'est plus une somme des parties qui le constitue, mais un tout, constitué de tous les aspects de la vie. Dans les comptes-rendus d'hospitalisation, la situation sociale, professionnelle ou encore les liens que le patient entretient avec son entourage sont décrits, pour mettre en lumière le contexte social, dans lequel évolue le patient.

Nous avons choisi de développer une modélisation ontologique sur les facteurs sociaux et environnementaux, afin de définir un cadre pour l'étude d'une possible corrélation entre les événements de la vie sociale et les troubles de la santé mentale.

---

7. La réalité clinique traduit le rapport entre théorie et pratique de la médecine. La théorie ne correspond pas toujours aux observations pratiques et inversement. C'est pourquoi les classifications en psychiatrie, bien que documentées et consensuelles dans leur usage, ne permettent pas toujours d'effectuer un codage diagnostique cohérent entre la réalité observée en suivi clinique et la réalité décrite dans une classification.

### **Proposer une interopérabilité des classifications psychiatriques utilisées au Service-Hospitalo Universitaire de Sainte-Anne**

Les professionnels en santé mentale disposent de différents systèmes de classification pour les aider dans leur diagnostic et il est rare que leur utilisation se restreigne à un seul de ces systèmes. À l'hôpital Sainte-Anne il est d'usage d'en utiliser trois différents. Et rien ne permet aux utilisateurs de ces classifications de faire formellement et automatiquement un lien entre les codages diagnostics. Pourtant, les politiques de codage économique obligent, du moins en France, à l'utilisation d'un système bien précis (la Classification Internationale des Maladies version 10 actuellement).

Nous avons décidé de développer un modèle qui permettra d'avoir accès aux catégories diagnostiques de trois grandes classifications psychiatriques utilisées au Service-Hospitalo Universitaire de l'hôpital Sainte-Anne.

### **Proposer un modèle de connaissances psychiatriques qui soit le résultat d'observations cliniques au Service-Hospitalo Universitaire de Sainte-Anne**

Nous souhaitons développer un modèle avec des termes extraits des CRH qui sont le reflet des observations cliniques faites par les praticiens. Ce module sera un ensemble organisé de concepts qui servent à décrire des situations cliniques.

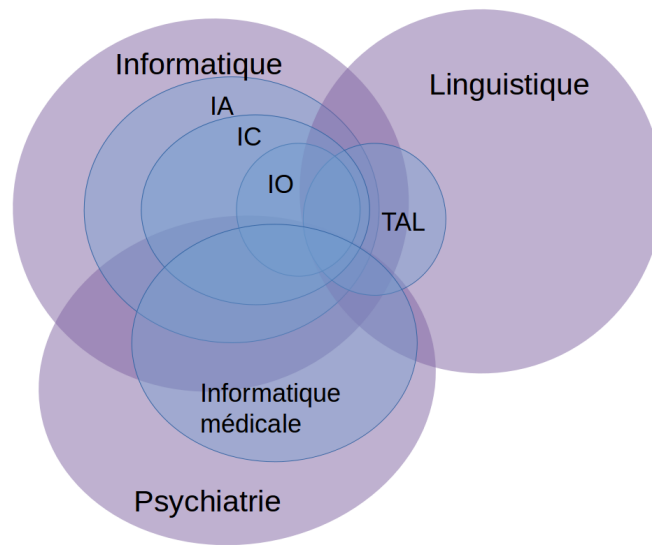
L'objectif est de proposer une alternative aux systèmes actuels de classification et de réduire l'incohérence entre la description des troubles et la réalité clinique. Ce modèle est à destination du Service-Hospitalo Universitaire de Sainte-Anne.

## **Contributions à la croisée de plusieurs disciplines**

Afin d'éclairer notre propos, nous détaillons nos contributions aux croisements disciplinaires de l'informatique, la linguistique et la psychiatrie selon le schéma 1 réalisé par Baneyx [2007].

**L'anonymisation d'entités nommées et l'étude comparative de deux systèmes d'extraction de termes candidats** est située dans le domaine du traitement automatique du langage (TAL). Le TAL est au croisement de la linguistique, l'informatique et l'intelligence artificielle (IA). Cette discipline vise à la description, la modélisation, ou encore la représentation du langage écrit ou parlé afin de : (1) reconnaître ou reproduire le langage (pour des outils de traduction ou de correction orthographique), (2) comprendre et interpréter le langage (pour des outils de reconnaissance vocale par exemple), (3) extraire de l'information contenue dans le langage (pour réaliser des outils d'extraction de termes spécialisés entre autre). Le TAL se base aussi bien sur la grammaire, la phonétique ou la sémantique des langues, que sur les statistiques, la logique ou les techniques d'apprentissage automatique pour le développement d'applications. Au cours de notre travail, nous avons réalisé l'anonymisation d'entités nommées à l'aide de l'outil MEDINA. Cette contribution est présentée au **chapitre 3**. Nous avons également réalisé l'étude comparative de deux systèmes d'extraction de termes candidats. Dans un premier temps, nous avons sélectionné YATEA, un outil utilisé pour identifier un groupe de mots pouvant correspondre à des termes spécialisés dans un texte Aubin et Hamon [2006]. Mais face à des résultats mitigés, nous avons

FIGURE 1 – Interdisciplinarité de nos contributions.



décidé de tester un nouvel extracteur paru en 2015, soit deux ans après le début de notre étude, BIOTEX. L'étude comparative de ces deux extracteurs est présentée au **chapitre 5**.

**Le développement des modules ontologiques** se situe dans le domaine de l'ingénierie ontologique (IO). Cette discipline vise au développement de méthodes pour représenter les connaissances dans des ontologies, et les exploiter dans des applications dédiées. Elle se base donc sur la linguistique, les langages informatiques de modélisation des connaissances, ou encore la logique informatique (le **chapitre 2** présente les principaux fondements de l'IO). Notre travail a contribué à la mise en œuvre des méthodes hybrides de développement d'ontologies. Ces contributions sont présentées dans les **chapitres 5 et 6**. En outre, nous nous sommes intéressés aux problématiques liées à la validation des ontologies (**chapitre 4**). Nous avons développé une méthode de validation structurelle et sémantique des ontologies, la méthode LOVMI. Elle repose sur des outils existants et une approche interactive avec les experts du domaine modélisé. Cette contribution est présentée au **chapitre 7**.

**La construction des modules de l'ontologie de la psychiatrie ONTOPSYCHIA** se place dans le domaine plus général de l'informatique médicale. Nous avons contribué au partage des connaissances en psychiatrie, par le biais de nos deux modules ontologiques. Nous proposons ONTOPSYCHIA, une ontologie pour la psychiatrie, composée de deux modules indépendants : (1) les facteurs sociaux et environnementaux des troubles mentaux, (2) les troubles mentaux. Associée à des outils dédiés, l'utilisation d'ONTOPSYCHIA permettra (1) d'effectuer des recherches sémantiques dans les CRH, (2) d'indexer les CRH pour la constitution des cohortes, (3) de représenter la comorbidité, (4) d'amorcer un consensus

autour des catégories descriptives des troubles mentaux, (5) d'étudier de possibles corrélations entre le contexte social et la santé mentale.

## Organisation du manuscrit

La **première partie** de ce manuscrit fait état des recherches dans les domaines connexes à notre étude.

**Le Chapitre 1** pose un certain nombre de questions pour présenter l'organisation et la représentation des connaissances. Ce premier chapitre nous permet de placer notre objet d'étude, l'ontologie informatique, dans son contexte général : la gestion de la connaissance.

**Dans le Chapitre 2** nous tentons de répondre aux questions liées aux ontologies. Les ontologies offrent la possibilité de développer des modèles conceptuels dans un formalisme unifié. Les chercheurs en biologie et en médecine se sont donc rapidement appropriés ces artefacts, pour intégrer l'ensemble des données hétérogènes et diverses dont ils disposaient et disposent encore. D'autres disciplines ont suivi cette démarche, pour organiser la sémantique de leurs données dans des modèles ontologiques. C'est ainsi que les ontologies sont devenues consensuelles, pour la représentation et la modélisation des connaissances. Ce deuxième chapitre permet de présenter les ontologies informatiques.

**Le Chapitre 3** s'intéresse aux méthodes de développement des ontologies : choix de la modélisation sémantique, acquisition des données sémantiques, ou encore choix des outils de construction. Le développement du modèle ontologique peut être étudié selon deux approches à la fois opposées et complémentaires : descendante (top-down) et ascendante (bottom-up). À cela s'ajoute les méthodes hybrides de plus en plus adaptées aux problématiques soulevées par l'automatisation du processus de développement, et l'accroissement d'ontologies disponibles et réutilisables. Dans ce troisième chapitre, nous explorons les différentes méthodes existantes pour le développement des ontologies.

**Dans le Chapitre 4** nous abordons la problématique majeure de la validation des conceptualisations mises en œuvre au sein des ontologies. Nous articulons notre propos autour de deux aspects complémentaires : (1) la *validation structurelle*, qui peut être réalisée automatiquement grâce au développement d'outils dédiés et (2) la *validation sémantique*, qui peine encore à trouver des méthodes consensuelles. Ce quatrième chapitre pose donc les bases de la validation des ontologies et de leurs modélisations.

La **deuxième partie** de ce manuscrit présente les travaux théoriques et pratiques réalisés au cours de cette thèse.

**Dans le Chapitre 5** nous présentons les différentes étapes de la conceptualisation du module sur les facteurs sociaux et environnementaux. Pour le développement du modèle,

nous avons suivi la méthodologie hybride TOREUSE2ONTO [Drame, 2014]. Pour développer la conceptualisation du domaine nous avons suivi les engagements ontologiques de la méthode ARCHONTE Bachimont *et al.* [2002], et en particulier le principe de normalisation sémantique.

**Dans le Chapitre 6** nous présentons la construction de la deuxième partie d'ONTOPSYCHIA. Nous avons réalisé un alignement de classifications pour construire un premier modèle. Ensuite, nous avons construit un deuxième modèle par méthode ascendante, à partir de l'extraction de termes candidats extraits des CRH de l'hôpital Sainte-Anne.

**Enfin, le Chapitre 7** répond aux questions liées aux outils qui permettent de valider la structure de notre ontologie. Nous présentons nos travaux qui ont mené à la proposition de la méthode LOVMI pour LES ONTOLOGIES VALIDÉES PAR MÉTHODE INTERACTIVE. Nous avons expérimenté cette méthode sur le module des facteurs sociaux et environnementaux d'ONTOPSYCHIA

## **Première partie**

### **État de l'art**



# Chapitre 1

## L'organisation des connaissances

### Sommaire

<b>1.1 Les systèmes d'organisation de la connaissance (SOC)</b>	<b>24</b>
1.1.1 Définitions	24
1.1.2 Présentation des SOC	25
<b>1.2 L'approche de la sémiologie et de la sémiotique pour la modélisation des connaissances</b>	<b>27</b>
1.2.1 Définition générale de la sémiotique	27
1.2.2 Définition du signe linguistique	28
1.2.3 Les grandes notions de la sémiotique :	29
1.2.4 Application de la théorie sémiotique à la médecine	31
<b>1.3 Un point sur les classifications utilisées en psychiatrie</b>	<b>31</b>
1.3.1 SNOMED 3.5VF and SNOMED CT	32
1.3.2 Classification statistique Internationale des Maladies et des problèmes de santé connexes	33
1.3.3 Diagnostic and Statistical Manual of Mental Disorders	34
1.3.4 Classification Française des Troubles Mentaux de l'Enfant et de l'Adolescent	34
1.3.5 Le Research Domain Criteria	35
<b>1.4 Synthèse</b>	<b>38</b>



*La connaissance, immatérielle, se transmet au travers de l'écrit, le dessin, le chant, la parole, la vidéo, ou encore la photographie. Elle se justifie au travers de l'expérience, de la recherche ou du raisonnement. Quand elle est modélisée dans des systèmes informatiques elle peut apprendre sur elle-même et produire de nouvelles connaissances. Quels sont ces systèmes qui organisent la connaissance ? Comment peut-on modéliser la connaissance dans des systèmes formels ? Et à quoi nous sert-il d'organiser cette connaissance ? Ces questions ont façonné ce premier chapitre qui présente l'organisation et la représentation des connaissances. Il nous permet de placer notre objet d'étude, l'ontologie informatique, dans son contexte général : la gestion de la connaissance. Nous nous sommes intéressés en particulier aux théories linguistiques, qui offrent un cadre de réflexion autour de la signification des concepts utilisés dans le langage. Dans les deux premières sections, nous centrons notre propos sur les systèmes d'organisation de la connaissance et sur l'apport de la sémiotique, pour la modélisation de la connaissance. Dans la troisième section, nous présentons l'organisation de la connaissances en psychiatrie, dans des classifications de référence.*

## 1.1 Les systèmes d'organisation de la connaissance (SOC)

### 1.1.1 Définitions

Le terme générique de système d'organisation des connaissances mêle des notions telles que organisation de l'information, structure sémantique d'un domaine, connaissances en interaction, ou encore les notions de but visé, et de point de vue sur le monde. Nous avons sélectionné trois définitions qui, selon nous, définissent efficacement ce que sont les SOC :

**Définition de Hodge [2000] :** « The term knowledge organization systems is intended to encompass all types of schemes for organizing information and promoting knowledge managements. knowledge organization systems include classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, and authority files that control variant versions of key information (such as geographic names and personal names). They also include less-traditional schemes, such as semantic networks and ontologies. »<sup>1</sup>. Hodge [2000] identifie les caractéristiques communes à tout SOC :

1. Un SOC impose un certain point de vu exprimé sur le monde par les collections et les items qu'il contient.
2. Une même entité peut être caractérisée de manière différente selon le SOC dans lequel elle est organisée.
3. La ressemblance entre un concept et l'objet du monde réel auquel il réfère doit être adéquate. Le système doit pouvoir être utilisé avec confiance et garantir l'interopérabilité dans son utilisation. En effet, une personne qui souhaite faire une recherche

1. « Le terme « systèmes d'organisation de la connaissance » vise à englober tous les dispositifs pour organiser l'information et promouvoir la gestion des connaissances. Les systèmes d'organisation de la connaissance incluent des dispositifs de classification qui organisent des documents à un niveau général (tels que des livres sur une étagère), des rubriques qui fournissent un accès plus détaillé et des fichiers d'autorité qui contrôlent les versions des informations clés (telles que les noms géographiques et les noms de personne). Ils incluent également des dispositifs moins traditionnels, tels que les réseaux sémantiques et les ontologies. »

## 1.1 Les systèmes d'organisation de la connaissance (SOC)

---

à l'aide d'un SOC doit pouvoir connecter son concept (et sa représentation) à un concept déjà présent dans le système.

**Définition de Zeng et al. [2007] :** « Ces systèmes modélisent la structure sémantique sous-jacente d'un domaine et fournissent des éléments sémantiques, de navigation et de traduction à l'aide d'étiquettes, de définitions, de typologies, de relations et de propriétés des concepts. »

**Définition de Vandenbussche [2011] :** « Un SOC est un ensemble de connaissances en interaction, représentées et regroupées au sein d'une structure dans le but de répondre à des besoins et d'atteindre des objectifs déterminés. »

### 1.1.2 Présentation des SOC

La volonté de représenter la connaissance au sein de systèmes structurés n'est pas une nouveauté, ce qui explique la multiplicité de ces systèmes : liste, glossaire, vocabulaire contrôlé, thésaurus, taxinomie, sont quelques-uns des plus courants. Les frontières entre ces systèmes sont parfois minces, selon leur niveau de détail et de formalisme.

**Glossaire :** « Ouvrage de référence contenant une liste de mots - généralement dans l'ordre alphabétique - qui informe sur la prononciation, la forme, l'étymologie, la grammaire et le sens. Un dictionnaire bilingue est une liste de mots d'une langue dans l'ordre alphabétique, accompagné de leur sens et de leurs équivalences dans une autre langue donnée. » [MeSH, 2008]

**Vocabulaire contrôlé :** « Liste de termes spécifiques au sens fixé et inaltérable, au sein desquels une sélection est faite pour le catalogage, l'analyse et l'indexation ou pour la recherche de livres, de périodiques ou d'autres documents. » [Prytherch et Bloomberg-Rissman, 1996]. Un vocabulaire contrôlé est donc exempt de l'ambiguïté du langage naturel et garant d'une uniformité lexicale dans la description des termes. Toutefois, lorsque cette liste de termes est organisée et régie par des relations sémantiques, nous parlons alors de thésaurus.

**Thésaurus :** « Ensemble de termes normalisés fondé sur une structuration hiérarchisée. Les termes y sont organisés de manière alphabétique et conceptuelle et reliés entre eux par des relations sémantiques. Un thésaurus forme un répertoire de termes normalisés pour l'analyse de contenu, le classement, l'indexation de documents d'information. » [Charlet et al., 2004]. Le MESH (Medical Subject Headings) fait office de thésaurus de référence dans le domaine biomédical. Il est utilisé pour l'indexation et l'interrogation de la base de données MEDLINE/PubMed. Dans sa version française, maintenue par l'INSERM, il est intégré au projet de création d'un Catalogage et d'une Indexation des Sites Médicaux de langue Française (le CiSMéF) [Darmoni et Joubert, 2000].

**Terminologie :** « Ensemble des termes particuliers à une science, à un art, à un domaine. Les termes y sont également définis par un texte en langue naturelle et caractérisés par différentes propriétés linguistiques ou grammaticales suivant l'usage prévu de cette terminologie. Avec leur mise sur support informatique, les terminologies ont beaucoup évolué et sont parfois enrichies de relations entre termes, formant ainsi un réseau terminologique. » [Charlet *et al.*, 2004]. Nous pouvons citer en exemple le dictionnaire Larousse Médical<sup>2</sup>, ou tout autre dictionnaire propre à un domaine donné.

**Classification :** Elle est le résultat de « l'action de distribuer par classes, par catégories » [Charlet *et al.*, 2004]. La classification n'inclue pas la définition des termes, l'importance est mise sur la distribution des termes dans des classes qui les définissent. L'ensemble des instances en œuvre dans une classification est appelé nomenclature.

**Classification à facette :** Ce système a été créé en 1924 par un bibliothécaire et mathématicien indien, Shiyali Ramamrita Ranganathan. Selon sa théorie, n'importe quel sujet peut être décrit à l'aide de cinq facettes : Personnalité, Matière, Énergie, Espace, Temps. Les facettes permettent donc de classer les documents selon leur sujet [Desfriches Doria, 2013]. L'avantage majeur de ce type de classification est de permettre la recherche documentaire selon des axes précis.

**Ontologie informatique :** « is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what « exists » is exactly that which can be represented. »<sup>3</sup> [Gruber *et al.*, 1993]. Une ontologie est un modèle de connaissances structurées qui permet de recenser, organiser et lier des concepts entre eux grâce aux relations qui les unissent. Les concepts sont des entités ayant un sens dans le domaine modélisé. Les relations peuvent être de nature sémantique, de composition ou d'héritage. Les ontologies sont des artefacts qui représentent un ensemble de connaissances compréhensibles et utilisables par une machine. BioTop<sup>4</sup> est un exemple d'ontologie sur le domaine de la biomédecine.

---

2. <http://www.larousse.fr/archives/medical>

3. « Est une spécification explicite d'une conceptualisation. Le terme est emprunté à la philosophie, où une ontologie est une prise en compte systématique de l'Existence. Pour les systèmes à base de connaissances, ce qui « existe » est exactement ce qui peut être représenté ».

4. <https://biportal.bioontology.org/ontologies/BT>

## 1.2 L'approche de la sémiologie et de la sémiotique pour la modélisation des connaissances

TABEAU 1.1 – Présentation sommaire des différents SOC définis dans ce chapitre.

SOC	Description	Exemple
<b>Glossaire</b>	Liste de mots dans l'ordre alphabétique	Dictionnaire bilingue
<b>Vocabulaire contrôlé</b>	Liste de termes spécifiques	Les vocabulaires contrôlés en œuvre dans les systèmes d'indexation
<b>Thésaurus</b>	Ensemble de termes normalisés, hiérarchisés et reliés par des relations sémantiques	Les thésaurus misent en œuvre dans les systèmes de classement de documents
<b>Terminologie</b>	Ensemble des termes particuliers à une science, à un art, à un domaine, accompagnés de leur définition en langage naturel	Dictionnaire spécialisé
<b>Classification</b>	Ensemble de mots regroupés dans des classes, des catégories	Les classifications médicales
<b>Classification à facette</b>	Ensemble de termes regroupés dans des classes elles-mêmes décrites par des facettes	Classification de la catégorie <i>lieux</i> selon les facettes <i>ville, pays, capitale, hameaux, lieux-dits</i> décrivant chaque terme [Desfriches Doria, 2013].
<b>Ontologie informatique</b>	Modèle de connaissances structurées qui permet de recenser, organiser et lier des concepts entre eux grâce aux relations qui les unissent	Les ontologies relatives au domaine médicale

## 1.2 L'approche de la sémiologie et de la sémiotique pour la modélisation des connaissances

### 1.2.1 Définition générale de la sémiotique

L'utilisation du terme sémiotique tend à se généraliser, pour désigner la science qui étudie les signes. Cependant, à l'origine de cette science, à la fin du 20<sup>ème</sup> siècle, nous distinguons deux approches différentes. La sémiologie tout d'abord, qui a fait son apparition avec les travaux de Ferdinand de Saussure. Il l'a défini dans son Cours de linguistique générale [De Saussure, 1989] comme « la science générale de tous les systèmes de signes (ou de symboles) grâce auxquels les hommes communiquent entre eux ». Cette approche est qualifiée de sociale, car elle vise à mettre en évidence l'organisation du langage en tant que système de communication. En parallèle aux travaux de Saussure, le terme sémiotique est apparu dans les travaux de Peirce [Peirce, 1978] pour désigner « une doctrine quasi nécessaire ou formelle des signes », « la science formelle des conditions de la vérité des représentations ». Cette approche est qualifiée par son auteur de « logique », elle vise à mettre en évidence les processus en œuvre dans la signification.

Les théories sémiotiques sont aujourd'hui au cœur de nos systèmes informatiques pour la gestion des connaissances sémantiques. Elles s'expriment au travers des concepts

que nous définissons dans les SOC, au travers des faits que nous modélisons dans les ontologies, au travers des relations qui nous permettent de lier des concepts entre eux. Nous centrons notre propos sur les grandes notions de la sémiotique et en particulier sur l'étude du signe linguistique, en lien avec les travaux des deux pères fondateurs de cette science, Ferdinand De Saussure et Charles Sanders Peirce.

### 1.2.2 Définition du signe linguistique

À la fin du 20<sup>ème</sup> siècle, **De Saussure** [1989] dans son Cours de linguistique générale, définissait le signe comme la combinaison d'un concept et d'une image acoustique, respectivement appelés le signifié et le signifiant (figure 1.1).

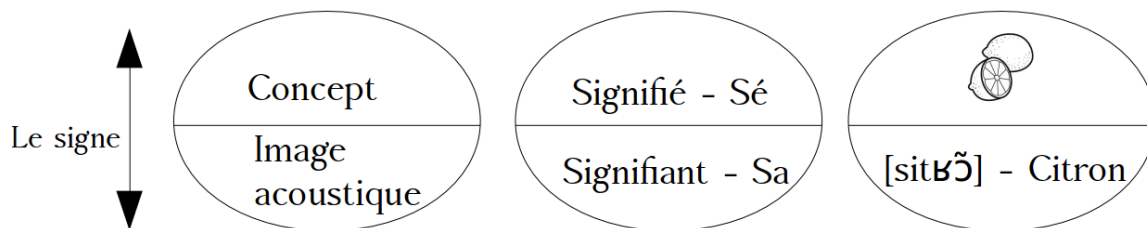


FIGURE 1.1 – Le signe linguistique selon Ferdinand de Saussure.

En parallèle à ces travaux, Charles Sanders Peirce définit le signe non pas sous la forme d'une dichotomie, mais sous la forme d'une triade (figure 1.2). Pour Peirce, toute pensée est signe. Sa sémiotique se base sur l'universalité de la perception [**Peirce**, 1978].

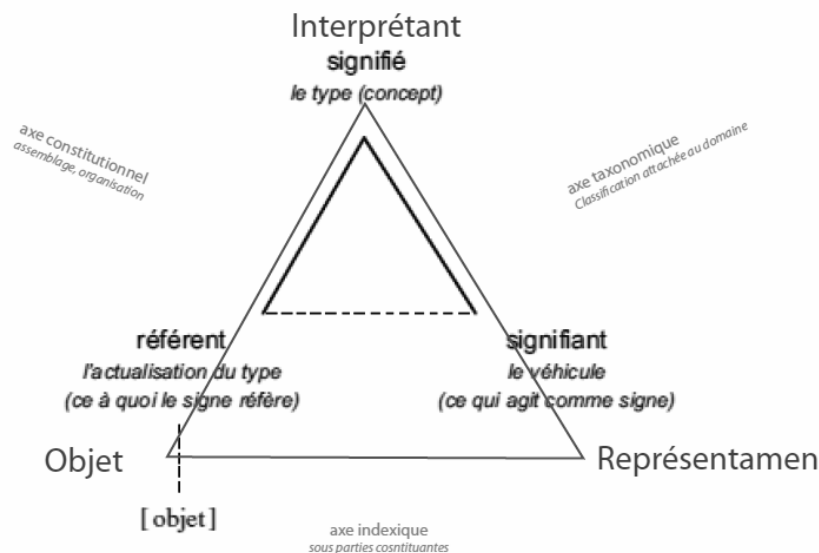


FIGURE 1.2 – Le signe linguistique selon Charles Sanders Peirce.

**Eco** [1984] s'inspire des travaux de Peirce pour développer sa théorie de la signification et définit formellement le signe comme suit : « le signe est utilisé pour transmettre une information, pour dire ou indiquer une chose que quelqu'un connaît et veut que les autres


connaissent également ». Il introduit la notion de « référent » à la dichotomie Sé/Sa de Saussure.

### 1.2.3 Les grandes notions de la sémiotique :

**Le signifiant** est l'image acoustique d'un mot. Il est la partie perceptible du signe. Les phonèmes qui composent un mot le différencient des autres et construisent alors les multiples signifiants. Le signifiant est donc sensoriel, il peut faire appel à l'ouïe, mais il peut faire appel à d'autres sens. Par exemple, le système des feux de circulation s'articule autour de trois couleurs (rouge, orange et vert) qui sont des signifiants visuels [Hébert et Dumont-Morin, 2012].

**Le signifié** est la partie intelligible du signe. Il porte le contenu sémantique associé au signifiant. Le signifié se décompose lui-même en sèmes qui le différencient des autres signifiés. Ainsi, la couleur verte du système des feux de circulation n'a pas le même signifié que la couleur rouge car il ne partage pas le même sème. Le premier dit « vous pouvez traverser », l'autre dit « arrêtez-vous » [Hébert et Dumont-Morin, 2012].

**Le référent** est ce que désigne le mot. L'objet du monde réel pointé par la combinaison du signifiant et de son signifié. Le dispositif du feu de circulation à une intersection routière est le référent pointé par la combinaison de la couleur de ces feux et de leur contenu sémantique [Hébert et Dumont-Morin, 2012].

**L'arbitraire du signe** est une notion importante de la sémiotique. La combinaison entre un signifiant et un signifié est dite arbitraire, car elle est immotivée. Il n'existe pas de rapport naturel entre le sens et sa réalisation visuelle ou acoustique, donc tout signifiant peut être associé à tout signifié. Dans le système des feux de circulation, les signes ont été codés pour que leur interprétation ne soit pas soumise à la subjectivité [Hébert et Dumont-Morin, 2012]. Si nous considérons des exemples du langage naturel, l'arbitraire du signe est plus évident. Par exemple, rien ne justifie que le signifiant [aʁbʁ] et le signifiant [tri:] désigne le même signifié [  ].

**La triade de Pierce** repose sur trois catégories philosophiques :

1. La priméité : « est le mode d'être de ce qui est tel qu'il est, positivement et sans référence à quoi que ce soit d'autre. » [Peirce, 1978]. Elle représente la qualité, l'universel et l'intemporel, la généralité dans l'ordre du possible. Peirce prend l'exemple du « rouge » ou plutôt d'un état d'être « rouge », indépendamment de tout autre chose. Pour que quelque chose puisse être « rouge » il faut qu'au préalable il existe du rouge en soi. « La priméité correspond à la vie émotionnelle » [Everaert-Desmedt, 2011].
2. La secondéité : « est le mode d'être de ce qui est tel qu'il est par rapport à un second, mais sans considération d'un troisième, quel qu'il soit. » [Peirce, 1978]. La secondéité est le lieu des faits, de l'individuel, de l'expérience, de l'existence, de l'action-réaction. La secondéité est toujours relative, car elle se définit par rapport au premier. « Une robe rouge » se définit par rapport à la priméité de l'existence du rouge. « La secondéité correspond à la vie pratique » [Everaert-Desmedt, 1990].

3. La tiercéité : « est le mode d'être de ce qui est tel qu'il est, en mettant en relation réciproque un second et un troisième. » [Peirce, 1978]. La Tiercéité représente la généralité dans l'ordre du nécessaire. Elle a valeur de prédiction, par exemple la loi de la pesanteur nous permet de prédire qu'à chaque pas que nous faisons nous revenons toujours sur le sol. Elle a valeur de loi, et règle les rapports entre les faits du second et les qualités du premier. « La tiercéité est la catégorie de la pensée, du langage, de la représentation, du processus sémiotique ; elle permet la communication sociale ; elle correspond à la vie intellectuelle. » [Everaert-Desmedt, 1990].

De ces trois concepts est née la définition du signe linguistique, qui s'articule autour de trois catégories [Everaert-Desmedt, 2011] :

1. Le representamen (proche du signifiant) : « il est une chose qui véhicule une autre chose ». C'est la catégorie de priméité.
2. L'interprétant (proche du signifié) : renvoie le representamen à son objet. Selon la théorie de la tiercéité, ce phénomène est en fait illimité. Pour interpréter un interprétant il nous faut à nouveau renvoyer un representamen à son objet. La définition d'un mot dans un dictionnaire en est un parfait exemple. Pour interpréter un representamen nous faisons appel à des interprétants qui eux mêmes renvoient le representamen à un objet, et ainsi de suite.
3. L'objet (proche du référent) : il représente le signe, en est l'objet. Peirce distingue (1) un objet dynamique, tel qu'il est dans la réalité et (2) un objet immédiat, tel que le signe le représente. Par exemple, dans « une robe rouge » la robe est l'objet dynamique et le rouge l'objet immédiat. L'objet dynamique (la robe) représente le representamen (la robe rouge) selon le point de vue de l'objet immédiat (le rouge).

Dans cette théorie, Peirce introduit la notion de pragmatique, dans le domaine de l'interprétant. Le sens en contexte est essentiel pour définir des concepts. Anna Wierzbicka ajoute à cela que la culture détermine également une part importante des concepts (signifiés) dénotés par les signifiants [Koselak, 2003]. Nous faisons également face dans les langues naturelles au phénomène de la polysémie. Ainsi, l'exercice d'interprétation du contexte d'apparition d'un signifiant est essentiel pour déterminer le signifié auquel il se réfère.

En annexe B nous avons reproduit le schéma d'illustration des théories de la sémiotique par Hébert et Dumont-Morin [2012]. Ce schéma a été réalisé en s'inspirant d'un texte de Rastier et de Eco. L'auteur a souhaité rapprocher les différentes appellations de chaque pointe du triangle sémiotique. Il insiste sur le fait que ce sont des rapprochements analogiques, et non des équivalences de termes. « Par exemple, l'interprétant de Peirce est, parmi les trois termes du signe tel que conçu par ce théoricien, ce qui se rapproche le plus de ce que Saussure appelle « signifié » ou de ce que Aristote appelle « états d'âme » ». L'auteur rappelle également que la signe linguistique selon son inventeur Saussure, n'est pas triadique, mais dyadique. Par ailleurs, la ligne qui relie le signifiant au référent est en pointillée, pour indiquer que cette relation est moins directe que les deux autres. Nous avons vu par exemple dans la théorie de Pierce, que le lien entre un representamen et un objet se fait exclusivement par l'interprétant.



### 1.3 Un point sur les classifications utilisées en psychiatrie

#### 1.2.4 Application de la théorie sémiotique à la médecine

##### Analyse sémique de Bernard Pottier

En médecine, la sémiologie fait référence à l'étude des symptômes et des signes, pour poser un diagnostic. Le but est de discriminer un référent (une maladie) selon les signifiés (symptômes) qui le composent. Dans les années 1960, Pottier élabore une méthode pour discriminer les concepts à travers les termes du langage [Pottier, 2001]. La méthode repose sur l'analyse sémique des signifiés. Dans une matrice, on représente des signifiés de même champ lexical qu'on discrimine par des traits sémantiques, des sèmes. Nous prenons l'exemple de l'analyse en sèmes de deux sous-types de la schizophrénie selon les critères du DSM IV TR (voir en section 1.3.3) :

TABEAU 1.2 – Analyse sémique de deux sous-types de la schizophrénie.

Sèmes	Schizophrénie	
	Type paranoïde	Type désorganisé
Idées délirantes	X	
Hallucinations	X	
Discours désorganisé		X
Comportement désorganisé		X
Symptômes négatifs	X	X

Dans le tableau, le signe X indique que le sème est présent pour le signifié.

Dans le tableau 1.2, nous observons que les deux sous-types de la schizophrénie ne partagent qu'un sème en commun « Symptômes négatifs » et sont discriminés par les quatre autres sèmes qu'ils ne partagent pas.

### 1.3 Un point sur les classifications utilisées en psychiatrie

Les classifications en psychiatrie sont dominées par deux courants de méthodologie antinomiques : la « pensée catégorielle » et l'« approche dimensionnelle » [Möller, 2008]. La « pensée catégorielle » domine les classifications actuelles en psychiatrie, mais est vivement critiquée depuis les années 1980 [Demazeux, 2008] au profit d'une « approche dimensionnelle ». La « pensée catégorielle » consiste à définir des catégories précises de troubles, de syndromes décrits par un ensemble de symptômes et de faits chronologiques. La version 10 de la Classification statistique Internationale des Maladies et des problèmes de santé connexes [WHO *et al.*, 1992] (CIM), ainsi que la 4<sup>ème</sup> édition du Diagnostic and Statistical Manual of Mental Disorders (DSM) appartiennent à cette catégorie de classification. Les opposants à cette pensée lui reprochent en particulier son inadéquation avec la réalité clinique et la catégorisation excessive doublée d'une trop grande rigidité des catégories diagnostiques [Demazeux, 2008, Widakowich *et al.*, 2013]. Ces difficultés posent la question de la délimitation des syndromes, dont les symptômes, bien souvent, appartiennent à plusieurs catégories diagnostiques. Le problème des « cas limites » est une parfaite illustration de la difficulté à poser un diagnostic via la classification catégorielle. En effet, certains patients présentent des symptômes qui peinent à entrer dans une



sous-catégorie rigoureuse. Afin qu'ils ne soient exempt des codages diagnostiques, une solution est d'avoir recours aux « NOS » *Not Otherwise Specified*, qui permettent d'attribuer une étiquette à un patient dont les symptômes ne permettent pas d'attribuer une sous-catégorie diagnostique définie.

Une seconde approche, dite « dimensionnelle » a émergé dans les années 1980 en réaction à la « pensée catégorielle » [Demazeux, 2008]. Les prémisses des classifications dimensionnelles sont visibles dans les travaux de Hempel [1965]. L'auteur met en avant l'intérêt non pas de classer des symptômes, mais d'ordonner des individus les uns par rapport aux autres selon des caractéristiques. Le but de cette approche est de faire apparaître « des distinctions plus subtiles que dans une classification » [Hempel, 1965, Demazeux, 2008]. Une telle approche cherche à mesurer des différences quantitatives d'un même trouble en établissant des degrés d'intensité dans les symptômes [Widakowich *et al.*, 2013]. Nous disposons pour cela d'outils d'évaluation clinique, tels que l'échelle de PANSS *Positive and Negative Syndrome Scale* pour la schizophrénie, ou l'IMC *Indice de Masse Corporel* pour les troubles alimentaires. Ces outils permettent de quantifier la sévérité d'un symptôme et non de définir de façon binaire, sa présence ou son absence [Widakowich *et al.*, 2013]. La dernière édition du DSM, la cinquième (voir 1.3.3) fait partie de ce type de classification, mais elle a été très mal reçue par la communauté scientifique, dont les critiques ont résonné dans la presse généraliste<sup>5</sup>. Allen Frances, le psychiatre responsable de l'édition du DSM-IV a publié un ouvrage critique sur la « médicalisation de la normalité » [Frances *et al.*, 2013] dans lequel il affirme que le « DSM-5 va convertir des millions de personnes normales en patients atteints de maladies mentales ». Ce fut ensuite au tour de l'Institut Américain de la Santé Mentale (National Institute of Mental Health, NIMH) de se désolidariser de cette édition et de poser par la même occasion la question de la pertinence des classifications catégorielles.

Dans la section suivante, nous présentons plus en détail différentes classifications relatives au domaine médicale et à la psychiatrie en particulier.

### 1.3.1 SNOMED 3.5VF and SNOMED CT

La première Systematized Nomenclature of Medicine (SNOMED) a été créée par le Dr Roger Cote en 1975. Ce système a évolué en (1) la SNOMED 3.5VF<sup>6</sup> en 1998, une terminologie multi axiale traitant des domaines de la médecine animale (y compris l'homme) et de la dentisterie humaine et (2) la SNOMED Clinical Term (CT)<sup>7</sup> en 2002, une ontologie qui représente une terminologie médicale clinique multilingue. La différence importante de ces classifications, par rapport au DSM ou à la CIM, est l'absence de règles ou de critères pour définir des catégories descriptives ou des symptômes.

La SNOMED3.5VF est détenue et distribuée par l'Agence des Systèmes d'Information Partagés de Santé (ASIP Santé)<sup>8</sup>, une agence d'État chargée de la e-santé en France. La SNOMED3.5VF attribue un code à tous les termes médicaux utilisés par les praticiens

5. [http://www.lemonde.fr/sciences/article/2013/05/13/dsm-5-le-manuel-qui-rend-fou\\_3176452\\_1650684.html](http://www.lemonde.fr/sciences/article/2013/05/13/dsm-5-le-manuel-qui-rend-fou_3176452_1650684.html), <http://www.healio.com/psychiatry/ptsd/news/online/>

6. <http://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/snomed-35vf>

7. <https://bioportal.bioontology.org/ontologies/SNOMEDCT>

8. <http://esante.gouv.fr/asip-sante>

en santé. Elle fonctionne comme un vocabulaire de santé contrôlé et unifié. Elle permet de stocker des informations médicales individuelles dans des entrepôts de données. Ce stockage vise à établir des outils d'analyse décisionnelle, à faciliter les décisions thérapeutiques, à contribuer aux études épidémiologiques et à l'enseignement. La première version française de la SNOMED a été réalisée en 1998. Des mises à jour de cette première version ont été effectuées depuis cette date, indépendamment des mises à jour de la SNOMED Internationale.

La SNOMED CT est détenue et distribuée par l'International Health Terminology Standards Development Organisation (IHTSDO)<sup>9</sup>. Les concepts sont uniques et représentent des aspects cliniques. Ils sont décrits en termes lisibles par l'homme et associés à une « balise sémantique ». Certains concepts sont liés à d'autres par des relations, qui fournissent également des définitions formelles ou des propriétés à ces concepts. La SNOMED CT vise à faire partie intégrante des applications pour soutenir l'information clinique. Tout comme la version française, elle fonctionne comme un vocabulaire contrôlé, une terminologie ou une classification, mais ne s'occupe pas des critères diagnostiques.

Les deux versions de la SNOMED contiennent un module qui ne traite que du contexte « social », qui vise à représenter les aspects sociaux qui peuvent influencer la santé et le traitement du patient.

#### 1.3.2 Classification statistique Internationale des Maladies et des problèmes de santé connexes

La classification de référence pour le codage médical dans les hôpitaux est actuellement la Classification statistique Internationale des Maladies et des problèmes de santé connexes [WHO *et al.*, 1992] (CIM) dont le Chapitre 05 « Troubles mentaux et du comportement » comporte plus de 1300 codes (extrait de la classification en annexe C). Elle est élaborée par l'Organisation Mondiale de la Santé (OMS) et principalement utilisée pour le codage du « Programme de médicalisation du système d'information » (PMSI). Le PMSI vise à introduire des concepts de comptabilité analytique dans la gestion administrative des hôpitaux : les diagnostics et actes effectués dans un établissement de santé sont codés et comptabilisés, rapportés à un patient et aux différents coûts de la structure. Cela permet de bâtir des indices de coûts relatifs par groupe homogène de malades. Le PMSI utilise un système de codage international, la CIM-10, pour les diagnostics, et un système français, développé grâce à une approche ontologique, la CCAM<sup>10</sup>, pour les actes. Le codage des diagnostics se fait en posant un diagnostic principal et, si nécessaire – au maximum 5 –, des diagnostics associés. Le PMSI a évolué vers une comptabilité qui vise à analyser le coût de chaque acte : c'est la tarification à l'activité ou T2A mais elle ne concerne pas la psychiatrie [Richard *et al.*, 2013]. Ce qui rend la CIM-10 populaire est également sa gratuité et sa facilité d'utilisation par des praticiens extérieurs à la psychiatrie (infirmières ou neurologues pour ne citer qu'eux), qui peuvent être impliqués dans le parcours de soin des patients. La CIM-10 comporte également un chapitre destiné à coder les facteurs environnementaux des maladies : le « Chapitre XXI : Facteurs influant sur l'état de santé et motifs de recours aux services de santé ». Ce chapitre est composé de sept sous groupes et d'un peu plus de 800 codes.

---

9. <http://www.ihtsdo.org/>

10. <http://www.ccam.sante.fr/>

### 1.3.3 Diagnostic and Statistical Manual of Mental Disorders

Le Diagnostic and Statistical Manual of Mental Disorders (DSM) [APA *et al.*, 2013] de l'Association Psychiatrique Américaine (APA) décrit et classe les troubles mentaux. L'APA indique sur son site internet <sup>11</sup> que le DSM est destiné aux milieux cliniques et aux cliniciens d'horizons théoriques différents. Il est à l'attention des professionnels des secteurs de santé mentale et autres, tel les psychiatres, les physicien(ne)s, psychologues, travailleur(e)s sociaux, infirmier(ère)s ou encore thérapeutes. La cinquième édition du DSM peut aussi être utilisée dans le cadre de recherches cliniques ou bien en tant qu'outil de collecte et de communication de statistiques sur la santé publique. Les trois composantes majeures du DSM sont :

- une classification des diagnostics : composée de la liste officielle des troubles mentaux reconnus par le DSM. Un code est associé à chaque diagnostic et utilisé pour le recueil de données ainsi qu'à des fins financières.
- un ensemble de critères associés à chaque diagnostic : indiquant les symptômes qui peuvent être présents ou liés à d'autres troubles.
- une description textuelle : qui accompagne chaque trouble répertorié, afin de fournir des informations concernant entre autres, les caractéristiques du diagnostic, le développement du trouble, les facteurs de risques, un diagnostic différentiel.

Le DSM est utilisé spécifiquement en France par les chercheurs cliniciens. Aux Etats-Unis, où il est plus populaire, il est utilisé aussi bien par les cliniciens que par les sociétés d'assurance, pharmaceutiques (pour la définition de critères de dosage thérapeutique et d'indications d'autorisation) ou par les pouvoirs publics et les dirigeants, pour les études de santé publique en particulier. Les utilisations du DSM sont également multiples, allant de l'aide au diagnostic, à la recherche, en passant par le codage médical. Le DSM est vivement critiqué en particulier dans sa version 5. Les détracteurs pointent par exemple, l'arbitrarité des catégories qui ne sont pas justifiées par des recherches étiologiques en biologie, des mécanismes neuronaux, des transmissions de maladies ou des prédispositions génétiques [Weinberger *et al.*, 2015]. Ces dernières années ont vu naître de nombreuses initiatives qui tentent de répondre aux manques et de combler les vides laissés par les catégories descriptives des troubles mentaux établies par l'APA.

### 1.3.4 Classification Française des Troubles Mentaux de l'Enfant et de l'Adolescent

La Classification Française des Troubles Mentaux de l'Enfant et de l'Adolescent (CFTMEA) établie sous la direction du Professeur Roger Misès a pour but de pallier certains manques dans le DSM ou la CIM sur les troubles propres à l'enfant ou à l'adolescent [Misés *et al.*, 2012]. La première édition de la CFTMEA a vu le jour dans les années 1980. Elle s'articule autour de deux axes : l'axe I des « catégories cliniques de base », et l'axe II des « facteurs antérieurs, éventuellement étiologiques ». La version de 2012 est alignée sur les codes de la CIM-10 pour faciliter le transcoding des patients. L'originalité de la CFTMEA est son approche classificatoire dimensionnelle inspirée de la psychanalyse. Le patient est un sujet qui possède une certaine structure psychique évolutive. Les diagnostics sont

---

11. <https://www.psychiatry.org/psychiatrists/practice/dsm>

donc une vue de cette organisation qui peut évoluer dans le temps et selon les interventions thérapeutiques. La classification est divisée en catégories principales qui « fixent la conduite à tenir et évaluent les risques à long terme » et de catégories complémentaires pour apporter des précisions au diagnostic. La CFTMEA en est actuellement à sa 5<sup>ème</sup> édition [Misès, 2012].

#### 1.3.5 Le Research Domain Criteria

Le projet Research Domain Criteria (RDoC) a débuté en 2009 sous l'égide de la National Institute of Mental Health (NIMH) et sous la direction de Bruce Cuthbert. Il est de loin le projet innovant le plus abouti en ce qui concerne la classification dimensionnelle en psychiatrie. Il offre un cadre de recherche pour étudier les maladies mentales sous un nouveau paradigme. Les classifications actuelles peinent à prendre en compte les avancées majeures réalisées dans les domaines de la génétiques, des neurosciences, de la cognition, et des maladies mentales en général depuis les années 1960 [Insel *et al.*, 2010, Weinberger *et al.*, 2015]. RDoC a donc pour ambition l'intégration de ces nouvelles connaissances pour mettre en lumière les fonctionnements de l'ensemble du comportement humain sur un axe allant de la normalité à l'anormalité. En outre, les auteurs de ce projet critiquent l'approche diagnostique actuelle, fondée uniquement sur un ensemble de symptômes. Ils questionnent également la validité de ces diagnostics en l'absence de tests ou de marqueurs biologiques justifiant les catégories descriptives des troubles mentaux [Insel, 2014].

**L'objectif** affiché dès le début du projet est le développement de nouvelles méthodes pour classer les troubles mentaux à des fins de recherche [Insel *et al.*, 2010, Morris et Cuthbert, 2012]. RDoC n'est donc pas à l'heure actuelle à considérer comme un nouveau système de classification complet et « prêt à l'emploi » [Cuthbert, 2014]. Morris et Cuthbert [2012] ajoutent qu'à long terme le projet veut permettre de (1) valider les tâches utilisées dans les essais cliniques, (2) identifier de nouvelles cibles pour le développement de traitements, (3) définir des sous-groupes cliniques significatifs en vue de la sélection des traitements, et (4) ouvrir la voie à des changements dans les décisions cliniques. Plus récemment, Insel [2014] et Cuthbert [2015] ont réaffirmé les objectifs du projet : « créer un système de classification expérimentale afin de faire un premier pas vers une médecine de précision<sup>12</sup> pour les troubles mentaux » [Cuthbert, 2015] ; « le but ultime de RDoC est la médecine de précision pour la psychiatrie, un système de diagnostic fondé sur une meilleure compréhension des bases biologiques et psychosociales d'un ensemble de troubles. » [Insel, 2014].

**Méthodes et résultats :** la matrice réalisée dans le cadre de ce projet, vise à l'intégration des caractéristiques génétiques, neuronales et comportementales dans la compréhension et la catégorisation des troubles. La matrice des RDoC à deux dimensions<sup>13</sup> est

---

12. La « médecine de précision », aussi appelée « médecine personnalisée » est un terme qui nous vient de l'oncologie. Elle a pour but de soigner les patients en fonction des caractéristiques de leur pathologie et des spécificités génétiques et environnementales qui peuvent différer d'un individu à l'autre.

13. Page du site du NIMH qui représente la matrice RDoC : <http://www.nimh.nih.gov/research-priorities/rdoc/constructs/rdoc-matrix.shtml>

composée des : (1) « RDoC Constructs » (les domaines, eux mêmes décomposés en sous domaines, présentés en annexe D) et des (2) « unités d'analyse ». Les « constructs » représentent l'organisation et le fonctionnement du cerveau. Les « unités (ou niveau) d'analyse » permettent de mesurer les « constructs ». L'intersection de ces deux axes est composée des recherches dans le domaine. Les domaines ont été définis par un groupe de travail du NIMH et les « constructs » et « sous-constructs » ont été définis entre 2010 et 2012 par des groupes de travail composés chacun d'environ 40 experts. Chaque groupe de travail devait réaliser trois tâches grâce à la littérature clinique [Cuthbert, 2014] : (1) définir les sous constructs, dans une liste préalablement établie par l'équipe RDoC du NIMH ; (2) fournir une définition à chaque sous construct ; (3) pour chaque sous construct, proposer des éléments caractéristiques aux unités d'analyses.

#### **Les 5 domaines principaux - constructs :**

1. Negative Valence Systems : « are primarily responsible for responses to aversive situations or context, such as fear, anxiety, and loss. » – Systèmes négatifs de valence : « sont principalement responsables des réponses aux situations ou au contexte aversif, comme la peur, l'anxiété et la perte. »
2. Positive Valence Systems : « are primarily responsible for responses to positive motivational situations or contexts, such as reward seeking, consummatory behavior, and reward/habit learning. » – Systèmes positifs de valence : « sont principalement responsables des réponses à des situations ou à des contextes positifs de motivation, tels que la recherche de récompense, le « consummatory behavior » et l'habitude de l'apprentissage de la récompense. »
3. Cognitive Systems : « are responsible for various cognitive processes. » – Systèmes cognitifs : « sont responsables des divers processus cognitifs. »
4. Social Processes : « mediate responses to interpersonal settings of various types, including perception and interpretation of others' actions. » – Processus sociaux : « réponses approfondies aux contextes interpersonnels de divers types, y compris la perception et l'interprétation des actions des autres. »
5. Arousal and Regulatory Systems : « are responsible for generating activation of neural systems as appropriate for various contexts, and providing appropriate homeostatic regulation of such systems as energy balance and sleep. » – Systèmes de stimulation et de régulation : « sont responsables de l'activation des systèmes neuroaux selon différents contextes et de la régulation homéostatique appropriée dans les systèmes tels que l'équilibre énergétique et le sommeil. »

#### **Les unités d'analyses :**

1. Gènes : par exemple l'« Acétylcholine » associée au sous-domaine « Menace Aiguë »
2. Molécules : par exemple la « Dopamine » associée au sous-domaine « Menace Aiguë »
3. Cellules : par exemple la GABAergic cells associée au sous-domaine « Menace Aiguë »

4. Circuits : permet d'associer des mesures de circuits neuronaux, étudiés au travers des techniques de la neuroimagerie, ainsi qu'au travers des modèles animaux ou de l'imagerie fonctionnelle.
5. Physiologie : permet d'associer des mesures bien établies, qui ont été validées lors de l'évaluation d'autres domaines, mais qui n'évaluent pas directement un circuit neuronal. Par exemple la fréquence cardiaque pour le sous-domaine « Menace aiguë ».
6. Comportement : permet d'associer des tâches comportementales ou d'observations du comportement à un domaine. Par exemple les expressions faciales dans l'étude de la « Menace aiguë ».
7. Auto-évaluation : permet d'associer des échelles d'entrevue, des questionnaires, ou tout autre instrument de mesure pour évaluer un domaine. Par exemple, le « Fear Questionnaire d'Isaac Marks » associé au sous-domaine « Menace aiguë ».
8. Paradigme : permet d'associer des tâches scientifiques significatives à un domaine. Par exemple les tâches sur la « peur conditionnée » pour mettre en évidence les circuits neuronaux impliqués dans les comportements de peur.

**Comment utiliser la matrice RDoC ?** Morris et Cuthbert [2012] rappellent que les systèmes de classification actuels imposent trois contraintes sur la variable indépendante des études psychiatriques : (1) les symptômes sont des unités d'analyse qui doivent être utilisées, (2) il faut utiliser une constellation particulière de symptômes (ceux présents dans le DSM ou leurs équivalents dans la CIM), (3) les symptômes sont utilisés de façon binaire (présent/absent) sans quantification. RDoC vise donc à libérer le chercheur de ces contraintes. Un exemple concret présenté dans Morris et Cuthbert [2012] : un patient présente un trouble lié à l'intériorisation (de l'humeur ou de l'anxiété). La matrice RDoC permet de classer les symptômes généraux du patient selon sa perception ou son évaluation<sup>14</sup> de la détresse ressentie (indépendamment des diagnostics du DSM). Ensuite il serait possible d'évaluer l'activation du circuit de la peur, afin d'établir un lien entre l'anxiété du patient et un sentiment de peur<sup>15</sup> dans une tâche adéquate (imagerie, vidéo). Afin de tester l'hypothèse suivante : la gravité et la chronicité de la détresse ressentie, par un patient atteint d'un trouble lié à l'intériorisation augmentent en réaction à l'hyporéactivité dans les circuits d'activation de la peur. Morris et Cuthbert [2012] pointent également le fait que les approches catégorielles qui analysent les symptômes de façon binaire (par exemple, en opposant présent à absent ou normal à anormal) sont incapables d'identifier des nuances dans l'affaiblissement cognitif des patients. Pourtant, les patients ne sont pas tous atteints par les symptômes de la même façon et selon la même intensité. Il apparaît de plus en plus important de pouvoir identifier la raison, la cause de ces nuances. Cuthbert [2014] propose également une méthodologie à l'attention des chercheurs pour

---

14. unité d'analyse : « Auto-évaluation »

15. unité d'analyse : « Circuit » ; domaine : « système à valence négative » ; « Negative Valence Systems are primarily responsible for responses to aversive situations or context, such as fear, anxiety, and loss » ; constructeur « Menace Aiguë » ; « Activation of the brain's defensive motivational system to promote behaviors that protect the organism from perceived danger. Normal fear involves a pattern of adaptive responses to conditioned or unconditioned threat stimuli (exteroceptive or interoceptive). Fear can involve internal representations and cognitive processing, and can be modulated by a variety of factors. »



passer du DSM/CIM à RDoC, dans le cadre de recherches cliniques, non pour poser un diagnostic.

**Les critiques** n'épargnent aucune nouvelle approche qui tend à sinon révolutionner, au moins apporter des changements significatifs dans un domaine donné. Phillips [2014] pressent que le message envoyé par l'initiative de la NIMH n'est rien d'autre qu'une proclamation de l'inutilité de l'Association Psychiatrique Américaine (APA) et de l'Organisation Mondiale de la Santé (OMS). L'auteur rappelle également que l'intérêt à court terme de la psychiatrie est l'accès à des soins de qualité pour l'ensemble des personnes atteintes de maladies mentales, peu importe leur revenu et leur situation sociale. L'auteur critique ainsi la vision à long terme du projet RDoC, qui ne se soucie guère des problématiques sociétales actuelles. L'auteur note également une forte confusion entre « circuit d'activation élevé » et « trouble », ce qui peut amener à des interprétations erronées. Par exemple, le circuit d'activation du sommeil est fortement élevé pendant le sommeil et pourtant ce phénomène n'est pas anormal. Il manque donc une dimension relative à l'évolution, pour pallier ce genre de confusion. L'auteur montre une légère amertume face à l'approche de RDoC concernant les facteurs de risque. En effet, RDoC tend à réorganiser les diagnostics en fonction des facteurs de risque partagés par des pathologies. Cependant, bien qu'un fumeur s'expose à un cancer et une maladie cardiovasculaire, ces deux maladies ne sont pas les mêmes. Enfin, nous pouvons citer Weinberger *et al.* [2015] qui questionne la validité des « domaines » et « unités », dont la majorité ne sont basés sur aucune expérimentation et dont le sens et l'utilité dans la réalité clinique sont pour l'heure inconnus. Pour aller plus loin, des études comparatives entre le DSM-5 et RDoC ont été publiées récemment, telles que Lilienfeld et Treadway [2016] ou encore Young [2016].

Pour conclure, à l'heure actuelle les RDoC n'ont pas pour vocation de remplacer le DSM ou la CIM dans le codage des maladies, mais de les supplanter dans les recherches cliniques [Yee *et al.*, 2015]. Afin de faire progresser la compréhension de la nature des troubles mentaux et de faire évoluer les catégories descriptives, à la lumière des neurosciences et de l'approche dimensionnelle. Par ailleurs, sur le long terme, le projet vise à atteindre l'exigence de la médecine de précision couplée à un système de classification des troubles mentaux, afin de proposer de meilleurs résultats dans les soins des pathologies psychiatriques [Insel, 2014, Cuthbert, 2015].

## 1.4 Synthèse

Dans ce chapitre nous nous sommes intéressés à la notion de concept et à ce que les concepts représentent en tant que connaissance sur le monde, au niveau intrinsèque (leurs sens dans le monde) et extrinsèque (leurs représentation dans le monde). Les systèmes d'organisation de la connaissance, qu'ils soient des classifications, des thésaurus ou des modèles conceptuels permettent d'intégrer et de modéliser ces concepts au sein d'une structure formelle, dont la puissance de représentation dépend du formalisme utilisé. L'organisation de cette connaissance permet alors de l'exploiter dans des applications dédiées. Par exemple, les classifications en psychiatrie offrent un cadre pour collecter la connaissance dans ce domaine. Elles sont ensuite utilisées pour poser des diag-

nostics, soigner des patients, faire des recherches cliniques ou encore coder des données médicales.

L'objet de nos travaux est le développement d'une ontologie informatique pour la modélisation de la psychiatrie. La difficulté majeure rencontrée par les praticiens en psychiatrie est l'absence de consensus autour des catégories descriptives des troubles psychiatriques, qui bien souvent ne reflètent pas la réalité qu'ils observent dans leur pratique clinique. L'analyse du sens des concepts telle que le propose la sémiotique semble pouvoir apporter une réponse méthodologique au développement des classifications.

Dans la suite de nos recherches, nous nous sommes concentrés sur les ontologies informatiques en tant que systèmes d'organisation de nos connaissances pour intégration dans un système à base de connaissances.





# Chapitre 2

## Système de représentation ontologique

### Sommaire

<b>2.1 L'ontologie informatique</b>	<b>42</b>
2.1.1 Définitions	42
2.1.2 Application, usage	43
2.1.3 Typologie des arborescences ontologiques	43
<b>2.2 Contexte : le Web Sémantique (WS)</b>	<b>45</b>
2.2.1 Naissance du Web 2.0	45
2.2.2 Langage du Web Sémantique	46
2.2.3 Les vocabulaires d'ontologies	53
<b>2.3 Les composants de la modélisation ontologique</b>	<b>55</b>
2.3.1 Notion de classes et d'instances de classe	55
2.3.2 Notion de propriétés, d'attributs, de rôles	56
<b>2.4 Synthèse</b>	<b>57</b>

*Dans le chapitre précédent, nous avons mis en avant l'importance des systèmes d'organisation et de modélisation des connaissances pour l'interaction entre les Hommes. Nous abordons dans ce chapitre l'aspect formel de la modélisation des connaissances, pour interagir au niveau machine. Nous nous intéressons en particulier aux ontologies informatiques et à leur puissance de représentation des connaissances. Ces artefacts permettent de regrouper des concepts, des relations entre ces concepts, des contraintes ou encore des règles. Les ontologies offrent la possibilité de développer des modèles conceptuels dans un formalisme unifié. Les chercheurs en biologie et en médecine se sont rapidement appropriés ces artefacts pour intégrer l'ensemble de données hétérogènes et diverses dont ils disposaient et disposent encore. De nombreuses disciplines ont suivi et organisé la sémantique de leurs données dans des modèles ontologiques. C'est ainsi que les ontologies sont devenues consensuelles pour la représentation et la modélisation des connaissances. La première section de ce chapitre est consacrée à la définition formelle de l'ontologie informatique et à la formulation des réponses applicatives qu'elle tend à apporter. Dans la deuxième section, nous faisons un point sur le Web Sémantique, qui a popularisé les ontologies et permis leur développement grâce à sa communauté. Enfin, la troisième section présente les composants d'une ontologie informatique.*

## **2.1 L'ontologie informatique**

### **2.1.1 Définitions**

Le mot ontologie est emprunté au mot latin scientifique « ontologia ». Ce dernier est composé du grec onto-, tiré du grec ancien ontos, qui signifie « étant, ce qui est », et de -logia, tiré du grec ancien logos qui signifie « discours, traité ». En philosophie, ce terme désigne une branche fondamentale de la métaphysique qui est selon Aristote « la science de l'être en tant qu'être ». L'ontologie s'occupe de ce qui existe, des propriétés générales de l'être. L'informatique a repris ce terme à la philosophie par analogie. En effet, les ontologies informatiques sont une représentation formelle de la connaissance, de ce qui existe dans le monde, et bien plus encore. Les définitions suivantes permettent de les décrire plus précisément :

**Définition de Gruber *et al.* [1993] :** une ontologie est la « spécification explicite d'une conceptualisation partagée pour un domaine de connaissance »

Cette définition a été l'une des plus reprises dans la littérature. Elle permet de préciser l'objet d'étude sans entrer dans des considérations d'appartenance à un courant du domaine. Aimé [2011] précise deux points qui émanent de cette définition : « la conceptualisation d'un domaine » se réfère à « un choix quant à la manière de décrire un domaine » ; et « la spécification de cette conceptualisation » est « sa description formelle ».

**Définition de Chandrasekaran *et al.* [1999] :** « Les ontologies représentent des théories sur différents objets, les propriétés de ces objets, et les relations entre ces objets qui sont acceptables dans un domaine de connaissance spécifique. »

Cette dernière citation permet d'expliciter plus précisément la portée d'une ontologie, en instaurant la notion d'objet et de relation entre ces objets. Cette notion d'objet sera reprise dans des définitions ultérieures. Ainsi, selon **Bourigault et al. [2004]**, une ontologie est « une conceptualisation des objets du domaine selon un certain point de vue ». Le point de vue de l'ontologie étant spécifié par les matériaux utilisés pour la construire. Le postulat de départ est que tout ce qui existe dans un système de connaissances peut être représenté. Ainsi, les noms des objets sont associés à du texte décrivant ce que ces noms désignent et des axiomes formels permettent de limiter l'interprétation que l'on peut faire de ces noms d'objets.

### 2.1.2 Application, usage

Les ontologies, qu'elles soient intégrées à des systèmes à base de connaissances (SBC) ou l'expression d'un vocabulaire commun propre à un domaine sont toujours développées avec un but applicatif précis, explicité au préalable. En s'appuyant sur les travaux de **Chandrasekaran et al. [1999]**, **Noy et al. [2001]** et **Charlet [2002]**, nous présentons un certain nombre de problèmes que peuvent résoudre les ontologies, sous deux axes complémentaires.

1. **Les ontologies en tant qu'artefacts intégrés à des SBC.** Elles clarifient la structure des connaissances, elles sont le cœur de tout SBC. **Charlet [2002]** parle de « squelette à la représentation des connaissances du domaine » et identifie plusieurs points qu'une ontologie intégrée à un SBC permet de résoudre :
  - L'interopérabilité.
  - L'indexation et la recherche d'information.
  - L'annotation conceptuelle
2. **Les ontologies en tant qu'objet de communication entre humains et avec les machines.** Elles permettent de partager des connaissances et un vocabulaire commun (ou standardisé) à un groupe de personne. L'ontologie a pour but de capturer l'essence des termes, le signifié, le concept et de leurs associer leur réalisation dans le langage courant, leur référent. En tant que tel, **Noy et al. [2001]** identifie plusieurs problèmes que permettent de résoudre les ontologies :
  - « Partager la compréhension commune de la structure de l'information entre les personnes ou les fabricants de logiciels »,
  - « Expliciter ce qui est considéré comme implicite sur un domaine »,
  - « Distinguer le savoir sur un domaine du savoir opérationnel »,
  - « Analyser le savoir sur un domaine ».

### 2.1.3 Typologie des arborescences ontologiques

Dans le cas d'ontologies sous formes d'arbre, les concepts se placent à un niveau précis de l'arborescence de l'ontologie. Un ensemble de concepts placés au même niveau de la hiérarchie aura un degré conceptuel différent de l'ensemble de concepts modélisés au

dessus ou en dessous dans l'arbre. Cette distinction de niveau conceptuel amène à distinguer différents types d'ontologies. La typologie la plus utilisée dans le domaine distingue trois niveaux dans les ontologies, relatifs à trois degrés d'abstraction : les top-ontologies, les ontologies génériques, les ontologies de domaine et de tâche [Guarino, 1997, Stenzhorn *et al.*, 2007]. Les ontologies de représentation sont également un type particulier d'ontologies, mais elles n'entrent pas dans la catégorisation topologique en degrés d'abstraction, nous les présentons donc à part.

TABEAU 2.1 – Présentation de la typologie adoptée dans ces travaux.

Type d'ontologies	Niveau d'abstraction	Exemple
Top-ontologies	concepts généraux relatifs au monde et non à un domaine particulier	BASIC FORMAL ONTOLOGY (BFO)
Ontologies génériques	concepts généraux relatifs à un domaine particulier	BIO TOP
Ontologies de domaine	concepts spécifiques à un domaine particulier ou à une tâche particulière	ONTOLURGENCE
Ontologies de représentation	concepts relatifs aux primitives logiques qui représentent l'ontologie	ONTOLINGUA

1. **Les top-ontologies** : aussi appelées upper-level ontologies ou ontologies fondationnelles (ce dernier terme ayant l'avantage de décrire le rôle de ces ontologies, et pas seulement la place dans le niveau conceptuel). Elles décrivent des connaissances de haut niveau, modélisent des concepts « généraux » sur le monde tel que le temps, l'espace ou l'action [Guarino, 1997, Stenzhorn *et al.*, 2007]. L'ensemble de ces concepts et leurs relations doivent pouvoir être utilisés dans toutes les disciplines, car l'ontologie fondationnelle ne contient pas de concept relatif à un domaine particulier [Guarino, 1997, Declerck *et al.*, 2012]. BASIC FORMAL ONTOLOGY (BFO) <sup>1</sup> est un exemple de top-ontologie qui est actuellement utilisée par environ 130 ontologies.
2. **Les ontologies génériques** : aussi appelées core-domain ontologies, top-domain ontologies ou ontologies noyaux. Elles sont le lien entre les top-ontologies et les ontologies de domaine [Stenzhorn *et al.*, 2007]. Elles contiennent ainsi des concepts généraux relatifs à un domaine. BIO TOP <sup>2</sup> est un exemple d'ontologie noyau, elles regroupent les concepts généraux relatifs au domaine de la biologie. Declerck *et al.* [2012] mentionnent qu'il peut être difficile de différencier « concepts généraux » et « concepts particuliers » d'un domaine particulier. Les auteurs prennent en exemple une ontologie noyau de la médecine. Elle modélisera les concepts généraux de maladie ou de symptôme, les concepts qui couvrent tous les sous-domaines de la médecine. Alors que l'ontologie de domaine modélisera les concepts propres à un sous-domaine de la médecine, telle que la psychiatrie (« schizophrénie » ou « trouble du comportement »).
3. **Les ontologies de domaine ou de tâche** : nous distinguons les ontologies qui décrivent les concepts spécifiques d'un domaine particulier, dans le but de décrire ce domaine (les ontologies de domaine) et les ontologies qui décrivent les concepts

1. <http://ifomis.uni-saarland.de/bfo/>

2. <https://biportal.bioontology.org/ontologies/BT>

utilisés pour réaliser une tâche, dans le but de décrire cette tâche (les ontologies de tâches) [Drame, 2014]. Ces ontologies vont servir les applications [Declerck *et al.*, 2012]. Guarino [1997] les place d'ailleurs toutes deux au même niveau conceptuel. Nous pouvons citer l'ontologie ONTOLURGENCE<sup>3</sup> pour la modélisation des urgences médicales.

4. **Les ontologies de représentation** : regroupent un ensemble de concepts relatifs aux primitives logiques qui représentent l'ontologie [Charlet *et al.*, 2004]. ONTOLINGUA de Gruber *et al.* [1993] ou ONTOCLEAN de Guarino et Welty [2009] (cf 4.2.2) sont des exemples d'ontologies dont le but est de décrire d'autres ontologies.

## 2.2 Contexte : le Web Sémantique (WS)

### 2.2.1 Naissance du Web 2.0

Une discussion sur les ontologies ne saurait se passer d'une présentation du Web Sémantique. La naissance des méthodes, standards et technologies mises en œuvre dans la construction d'ontologies a été possible grâce au développement de ce Web 2.0. C'est le W3C<sup>4</sup> qui a formellement défini les objectifs du WS, les langages permettant de les atteindre et le projet majeur de formatage des connaissances non structurées dans des ontologies, en vue d'une interopérabilité et d'une interprétation de ces connaissances par un ordinateur. L'idée du WS est formulée par Berners-Lee *et al.* [2001] et révolutionne le Web de l'hypertexte. Cependant, le WS implique que les connaissances disponibles sur le Web sous la forme de données soient liées entre elles. Et cette étape indispensable peine à se mettre en place. Seulement cinq ans après la proclamation du Web Sémantique, celui-ci est renommé Web de Données. Une nouvelle définition est apportée, afin de se concentrer sur l'interconnexion des données du Web et soutenir le développement de ce projet et des technologies qui en dépendent [Shadbolt *et al.*, 2006]. Un an plus tard, l'un des plus gros projets pour l'interconnexion des données ouvertes et liées (DOL) du Web est lancé entre l'Université libre de Berlin et l'Université de Leipzig, avec en collaboration OpenLink Software, DBPedia<sup>5</sup>. Le but de ce projet est de fournir une version structurée et normalisée en langage du Web Sémantique du contenu de Wikipédia<sup>6</sup>. Une version française de DBPédia est réalisée par SémanticPédia<sup>7</sup>, une plateforme de collaboration entre le Ministère de la culture et de la communication, l'Inria et Wikimedia France. DBPedia est une ontologie, qui utilise le format de données RDF et contient une couche sémantique sous OWL (voir en section 2.2.2). Par la suite, les gouvernements de différents états vont répondre à l'appel de Tim Berners Lee les invitant à mettre leurs données publiques à disposition du Web<sup>8</sup>. Le projet communautaire du « Linking Open Data » vise donc à publier sur le Web les données ouvertes (les données numériques libres d'accès et d'usage telles que les textes de lois, résultats d'élections, horaires de trains en temps réel, etc) pour ensuite mettre en relation ces données via le formalisme RDF.

---

3. <http://bioportal.bioontology.org/ontologies/ONTOLURGENCES?p=classes&conceptid=root>

4. Le W3C est un consortium qui travaille au développement des standards du Web <https://www.w3.org/>.

5. <http://wiki.dbpedia.org/>

6. <https://fr.wikipedia.org>

7. <http://www.semanticpedia.org/>

8. La France s'est positionnée en troisième place sur 184 en 2014. La plateforme [data.gouv.fr](http://data.gouv.fr) est dédiée au partage, à l'amélioration et à la réutilisation des données publiques.

En quelques années, le Web Sémantique a pris une ampleur considérable, grâce au projet communautaire pour les données liées, dont les membres sont très actifs. Nous pouvons observer cette ampleur sur les nuages qui représentent les liens entre les données ouvertes en annexe E ou sur le site du « Linking Open Data cloud diagram »<sup>9</sup>, pour visualiser les dernières mises à jour. Le Web Sémantique permet ainsi d'annoter chaque document référencé sur le Web, avec des métadonnées sémantiques qui pointent vers des éléments des données ouvertes, elles mêmes modélisées au sein d'ontologies [Stern, 2013]. Et ce système d'annotations reliées à une ontologie permet ensuite de faire des traitements automatiques sur les données. Cela prend forme notamment via les moteurs de recherche, qui se servent de ces annotations pour indexer leurs documents et permettre une recherche sur le sens des termes. Ainsi que dans les systèmes d'interaction Homme / Machine afin de faciliter entre autre le filtrage d'informations [Lenne, 2009].

Enfin, nous pouvons noter que, tout comme le Web de l'hypertexte, qui s'était développé sous l'impulsion des physiciens qui s'échangeaient massivement des documents, l'essor du Web Sémantique est directement lié au besoin d'intégration de données de disciplines majeures tels que la biologie et la médecine. De nombreuses disciplines ont suivi et organisé la sémantique de leurs données dans des modèles ontologiques. Tim Berners Lee participe également grandement à la popularité de cette technologie et porte, avec le W3C, une franche responsabilité dans l'avènement du Web 2.0.

### 2.2.2 Langage du Web Sémantique

Le Web Sémantique est une extension du Web, il en conserve donc les formalismes de données et vient ajouter en différentes couches supérieures ses propres formalismes. La figure 2.1 illustre ces technologies :

- Les URIs *Uniform Resource Identifiers* et les IRIs (*Internationalized Resource Identifiers* qui permettent d'identifier les ressources à l'aide d'une annotation unique. Les URIs sont définis sur l'Unicode, et les IRIs sont venues s'ajouter pour renforcer ce système d'identification des ressources du Web, en permettant à chacun d'utiliser sa propre langue pour identifier ses ressources (s'appuie sur la norme ISO 3166 qui associe un code à chaque pays).
- La couche relative aux formats de données est composée du langage à balise XML pour la création de documents structurés et interopérables, de Turtle RDFa (présenté en section 2.2.2).
- Pour l'échange d'informations, c'est le langage RDF (présenté en section 2.2.2) qui permet de définir et interconnecter les ressources/documents anonymes ou nommés par un URI/IRI [Héon, 2014].
- La représentation et la modélisation de l'information arrive au dessus (avec les langages RDFS présenté en section 2.2.2 et OWL présenté en section 2.2.2 et le vocabulaire SKOS présenté en section 2.2.3). Au même niveau, nous retrouvons les recommandations du W3C : RIF pour les langages de règles, SPARQL (présenté en section 2.2.2) pour les requêtes. Ces niveaux composent la partie logique du Web Sémantique.

---

9. <http://lod-cloud.net/versions/2017-02-20/lod.svg>

## 2.2 Contexte : le Web Sémantique (WS)

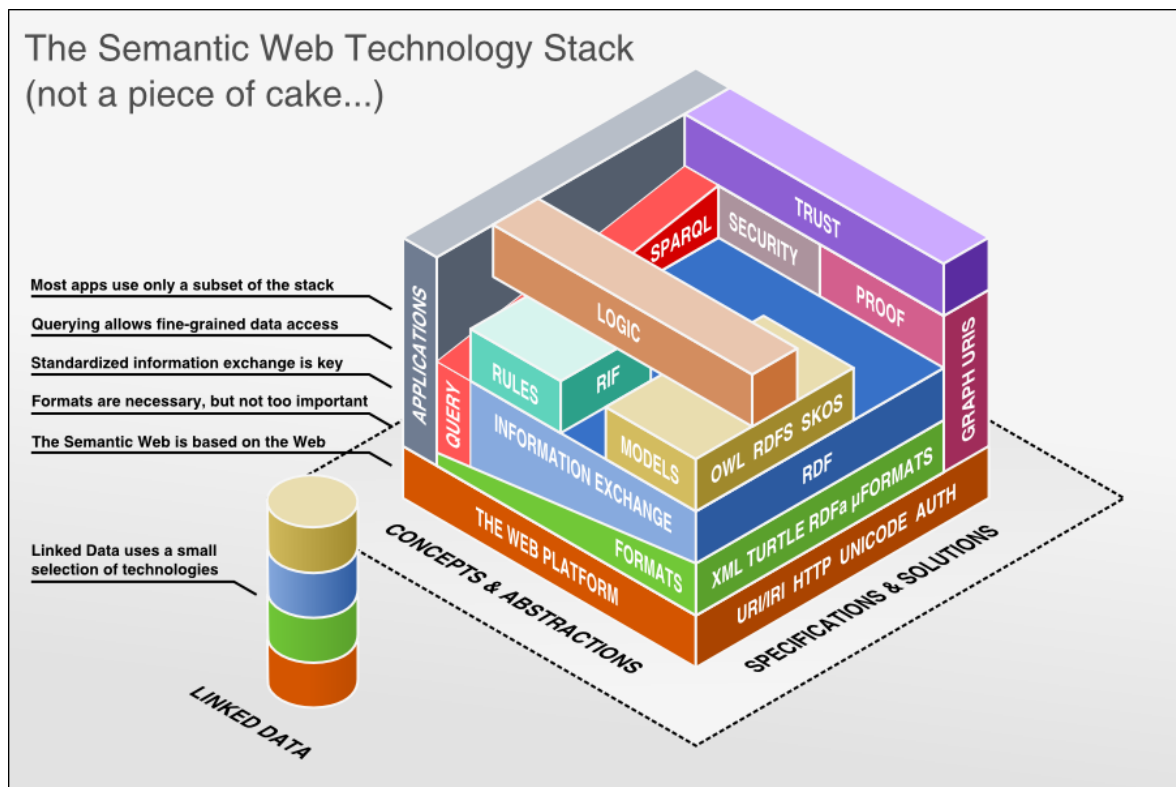


FIGURE 2.1 – The Semantic Web - Not a piece of cake...by Benjamin Nowack .

### Le langage Resource Description Framework (RDF)

Le formalisme RDF sert comme son nom l'indique de structure, de modèle de données, pour représenter les informations sur le Web [Smith *et al.*, 2004]. Le langage RDF modélise les données sous forme de graphe orienté. Un sommet sujet est relié à un sommet objet par un arc étiqueté. Nous parlons donc de relation prédicative entre un objet et un sujet.

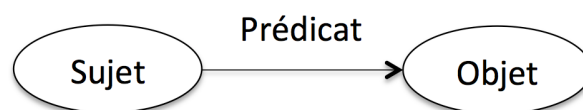


FIGURE 2.2 – Un graphe RDF représentant un triplet : deux nœuds (Sujet et Objet) reliés entre eux par une relation (le Prédicat) Smith *et al.* [2004].

Sur le Web, chaque ressource est identifiée par un URI unique. Le Web Sémantique se sert de ces URI pour lier les données entre elles (dans le Web des Données) qui sont elle-mêmes représentées dans une ontologie. Sur le Web, notre triplet RDF *La dépression est caractérisée par de la tristesse* sera donc modélisé comme tel :

- L'en tête du fichier RDF définit les espaces de noms utilisés dans le fichier : (1) l'espace de nom « rdf » contient l'adresse du fichier de syntaxe rdf ; (2) l'espace de nom



« TroubleMental » contient l'adresse de l'ontologie définie dans le document.

```
<rdf:RDF
  (1) xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  (2) xmlns:TroubleMental="http://www.semanticweb.org/TroubleMental#">
```

- Le corps du document RDF représente les données sous forme de graphe. Chaque ressource est identifiée par son URI : URL de l'espace de nommage (ici « <http://www.semanticweb.org/TroubleMental> », abrégé en « &TroubleMental ») + nom de la ressource (« dépression » ; « est\_caractérisé\_par » ; « tristesse »).  
(1) sujet ; (2) prédicat ; (3) Objet :

```
<rdf:Description (1) rdf:about="&TroubleMental;dépression">
  (2) <TroubleMental:est_caractérisé_par
  (3) rdf:resource="&TroubleMental;tristesse"/>
</rdf:Description>
```

### Le RDF Schema

Le RDF Schema est l'extension sémantique de RDF. Cette extension permet par exemple de regrouper des ressources dans des groupes. Elle permet donc de définir des classes, des propriétés et des instances, donc de définir le rôle des ressources selon leur utilisation.

- Exemple de la propriété « la dépression est caractérisée par un état maniaque » définie avec RDF Schema : (1) URI de la propriété ; (2) URI du domaine/sujet de la propriété ; (3) URI du rang/objet de la propriété.

```
<rdf:Description (1) rdf:about="&TroubleMental#est_caractérisé_par">
  (2) <rdfs:domain rdf:resource="&TroubleMental#Dépression"/>
  (3) <rdfs:range rdf:resource="&TroubleMental#Etat_Maniaque"/>
</rdf:Description>
```

- Exemple de la relation de subsomption « La dépression est une maladie » définie avec RDF Schema : (1) URI de la classe modélisée ; (2) URI de la classe mère de la classe modélisée.

```
<rdf:Description (1) rdf:about="&TroubleMental#Dépression">
  (2) <rdfs:subClassOf rdf:resource="&TroubleMental#Maladie"/>
</rdf:Description>
```

### Terse RDF Triple Language (Turtle)

Turtle est la syntaxe d'un langage qui permet une sérialisation non-XML des modèles RDF. C'est un sous-ensemble de la syntaxe Notation3 [Beckett et Berners-Lee, 2008]. En d'autres termes, ce langage permet d'écrire des graphes RDF sous la forme de texte libre, avec des abréviations. Nous reprenons ici notre triplet RDF *La dépression est caractérisée par de la tristesse* modélisé en Turtle :

- L'en tête du fichier Turtle définit les espaces de noms et leur URI utilisés dans le fichier : (1) « @prefix rdf » contient l'adresse du fichier de syntaxe rdf du document ; (2) @base fait référence à l'adresse de l'ontologie définie dans le document.

## 2.2 Contexte : le Web Sémantique (WS)

---

```
(1) @prefix rdf: <"http://www.w3.org/1999/02/22-rdf-syntax-ns#"> .  
(2) @base <"http://www.semanticweb.org/TroubleMental#"> .
```

- Le corps du document Turtle représente les données sous forme de graphe. Chaque nœud est identifié par son URI : URL de l'espace de nommage (ici « `http://www.semanticweb.org/TroubleMental` », abrégé en « `&TroubleMental` ») + nom du champ (« `dépression` » ; « `tristesse` » ; « `est_caractérisé_par` »). La différence avec la syntaxe RDF, c'est qu'il n'y pas de balise, pas d'indication typographique particulière indiquant le rôle des URI. Le triplet est représenté sous la forme d'une phrase : (1) sujet ; (2) prédicat ; (3) Objet :

```
(1) <&TroubleMental#dépression> (2) <&TroubleMental#est_caractérisé_par>  
(3) <&TroubleMental#tristesse> .
```

La syntaxe Turtle peut être couplée à d'autres langages d'ontologies afin d'en affiner la granularité sémantique.

### The OWL Web Ontology Language

Le RDFS n'offre pas une granularité très fine et permet de représenter une gamme très limitée de rôles sémantiques. Le W3C a donc proposé une nouvelle extension en 2003, le OWL [McGuinness et Van Harmelen, 2004]. Ce langage succède au DAML+OIL [McGuinness *et al.*, 2002] et s'appuie sur la logique de description (DL). Il permet de représenter formellement des connaissances, à un niveau de logique des prédicats. Il permet donc à une machine de raisonner sur une base de connaissances : inférer des connaissances implicites et détecter des incohérences. Il ajoute ainsi au RDFS des descriptions de relations entre classes plus complexes (intersection ou union par exemple), des propriétés ou des restrictions de cardinalités ou d'égalités entre autres [Laublet *et al.*, 2002, Reymonet *et al.*, 2009].

Exemple de la propriété « la dépression est caractérisée par de la tristesse » définie en OWL : (1) URI de la propriété ; (2) URI du domaine/sujet de la propriété ; (3) URI du co-domaine/objet de la propriété.

```
<owl:ObjectProperty (1) rdf:about="&TroubleMental#est_caractérisé_par">  
(2) <rdfs:domain rdf:resource="&TroubleMental#Dépression"/>  
(3) <rdfs:range rdf:resource="&TroubleMental#Tristesse"/>  
</owl:ObjectProperty>
```

On peut également préciser que OWL, dans sa première version, est constitué de trois sous-langages impliquant chacun un niveau de complexité croissant, chacun incluant son prédécesseur. Les spécificités de ces niveaux, ainsi qu'une aide au choix du niveau conceptuel est consultable dans le guide de référence de l'extension Protégé-OWL [Hortridge *et al.*, 2004] :

- **OWL Lite** : est utilisé uniquement pour la construction de hiérarchies simples tels que les taxonomies, avec la relation de subsumption. Il ne supporte pas l'intégralité des constructeurs de OWL tel que la négation ou la disjonction.
- **OWL DL** : se base sur la logique de description. Il est plus expressif que OWL Lite et supporte le raisonnement automatisé. Et à la différence de OWL Full, il garantit la complétude des raisonnements (les inférences/déductions sont toutes calculables) et leur décidabilité (leur calcul est réalisé en un temps fini).

- **OWL Full** : correspond au plus haut niveau d'expressivité. Il est utilisé pour décrire très finement des situations pour lesquelles la complétude et la décidabilité n'importent pas. Car, par son niveau de complexité, aucun raisonneur ne peut supporter un raisonnement complet sur une ontologie en OWL Full (par exemple un élément de l'ontologie peut être concept et instance).

## The OWL Web Ontology Language 2

Une révision de OWL est sortie en 2009 suite à un retour d'expérience des utilisateurs et utilisations du langage. Il inclut les sous langages Lite et DL de la première version de OWL. Les ontologies développées en OWL Lite ou OWL DL sont donc des ontologies valides en OWL 2 et ces deux sous langages de OWL 1 sont donc également des sous langages de OWL 2. Il permet toutefois une plus grande expressivité et de définir entre autres des classes disjointes (aucun individu ne peut être instance de deux classes), de faire des assertions négatives (not) ou encore d'étendre une restriction de cardinalité. Il est divisé en trois sous langage : QL, EL, RL (en plus de OWL Lite et OWL DL) [Group, 2008, Hitzler *et al.*, 2009].

**OWL 2 EL (Existential Language)** (ressemble au OWL DL) s'adresse principalement aux ontologies légères, qui contiennent un grand nombre de propriétés et/ou de classes avec une structure complexe. Il a été défini en suivant les exemples d'ontologies du domaine biomédical et il suffit à définir des ontologies telles que la SNOMED CT.

Exemple d'implémentation en OWL EL de la proposition « Une personne narcissique à pour objet d'amour elle même » avec l'utilisation de la restriction Self qui permet la réflexivité locale [Group, 2008] :

```
EquivalentClasses(
  :NarcisticPerson
  ObjectHasSelf( :loves ) )
```

**OWL 2 QL (Query Language)** s'adresse principalement aux ontologies qui contiennent un très grand nombre d'instances et pour lesquelles le plus important est la tâche de raisonnement liée au requêtage. L'expressivité de ce sous langage est donc réduite.

Exemple d'implémentation en OWL QL de la proposition « Toute personne sans enfant est parent d'aucune autre personne » [Group, 2008] :

```
SubClassOf(
  :ChildlessPerson
  ObjectIntersectionOf(
    :Person
    ObjectComplementOf(
      ObjectSomeValuesFrom(
        ObjectInverseOf( :hasParent )
        OWL:Thing ) ) ) )
```

**OWL 2 RL (Rule Language)** s'adresse principalement aux ontologies pour lesquelles le raisonnement évolutif sans perte de puissance d'expressivité<sup>10</sup> est le point important. Ce sous langage est le plus intéressant quand il est nécessaire de combiner le langage OWL avec des règles, trouver un compromis entre calculabilité et représentation [Héon, 2014]. Exemple d'implémentation en OWL RL de la proposition « Mary, Bill et Meg sont des femmes avec au plus une fille » [Group, 2008] :

```
SubClassOf(  
  ObjectIntersectionOf(  
    ObjectOneOf( :Mary :Bill :Meg )  
    :Female )  
  ObjectIntersectionOf(  
    :Parent  
    ObjectMaxCardinality( 1 :hasChild )  
    ObjectAllValuesFrom( :hasChild :Female ) ) )
```

### SPARQL

Les prémisses du langage SPARQL ont vu le jour en 2005, mais c'est en 2008 que le W3C en publie une première recommandation [Prud'Hommeaux *et al.*, 2008], avant une mise à jour en 2013 de la version 1.1. Le langage a été développé par le groupe de travail *W3C RDF Data Access Working Group (DAWG)*, aujourd'hui le *SPARQL Working Group*<sup>11</sup> [Harris *et al.*, 2013]. Pour rappel, RDF permet la représentation des informations sur le Web sous la forme de graphes orientés et étiquetés (voir la section 2.2.2 pour plus de détails). SPARQL est donc défini en suivant ce format et permet de rechercher des motifs de graphe (graph-matching). Les interrogations peuvent être faites sur des sources de données RDF ou sur des sources vues comme du RDF via un logiciel médiateur (middleware) [Prud'Hommeaux *et al.*, 2008, Harris *et al.*, 2013]. Le SPARQL est considéré comme une technologie clé du Web Sémantique, car il permet d'interroger l'ensemble des données liées au format RDF. Les requêtes en SPARQL sont composées de trois parties [Pérez *et al.*, 2009] :

1. Le filtrage par motif : il contient le motif de graphe, soit le motif de la requête, la clause de la requête. Il est contenu après le mot clé *WHERE*. Nous pouvons ajouter à ce motif : (1) des conditions à l'aide de *FILTER*, qui permet également de tester la non-existence d'un motif *FILTER NOT EXIST* ; (2) l'union du résultat de plusieurs motifs avec *UNION* ; (3) la possibilité de choisir la source de donnée à filtrer dans la requête avec *FROM*.
2. Le transformateur ou modificateur de solution : qui permet de modifier la sortie de la requête à l'aide d'opérateurs, tels que *DISTINCT* (pour éliminer les doublons), *ORDER* (pour ordonner les résultats), *LIMIT* (pour limiter le nombre de résultats), *OFFSET* (pour afficher les résultats à partir d'un nombre donné).
3. Le résultat de la requête SPARQL, qui peut avoir différents formats : *SELECT* pour la sélection de valeurs de variables qui vérifie le motif et restitue les résultats sous

---

10. L'expressivité, tel que pour le langage naturel, définit la capacité d'un langage ontologique à représenter une situation du monde réel ou une situation discursive. Il est possible d'exprimer en OWL un fait précis « Les Hommes sont mortels », mais il est beaucoup plus difficile d'exprimer une situation temporelle « Après la pluie vient le beau temps ».

11. [https://www.w3.org/2009/sparql/wiki/Main\\_Page](https://www.w3.org/2009/sparql/wiki/Main_Page)

forme de liste, *ASK* pour un résultat booléen, cette requête permet un test de vacuité, *DESCRIBE* pour l'obtention d'informations sur une ressource, *CONSTRUCT* pour la construction d'un nouvel ensemble de données RDF qui décrit les ressources résultats.

### **Exemple de requêtes SPARQL simple :**

En en-tête de la requête doivent être précisés les espaces de noms, via les *PREFIX*. Nous utilisons des variables suivi d'un « ? » pour récupérer les informations souhaitées dans les triplets, elles correspondent à tout nœud présent dans la ressource RDF.

### **Extraire les concepts associés à un code CIM-10**

```
(1) PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
(2) PREFIX ont: <http://www.limics.fr/OntoPsychia_Classifications#>

(3) SELECT ?concept ?labelCIM10
(4) WHERE { ?concept ont:CIM_10Id ?labelCIM10.}
```

(1) et (2) espace de noms pour le modèle RDFS et le vocabulaire de l'ontologie ; (3) sélection des variables à afficher en résultat ; (4) motif de la requête : nous cherchons dans les concepts ceux qui ont une annotation CIM\_10Id qui correspond à l'alignement du concept avec un code de la CIM-10. En résultat, nous affichons les concepts associés à un code CIM-10 et leur libellé.

### **Compter le nombre de concepts associés à un code CIM-10**

```
(1) PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
(2) PREFIX ont: <http://www.limics.fr/OntoPsychia_Classifications#>

(3) SELECT (COUNT(?labelCIM10) AS ?Concept)
(4) WHERE { ?concept ont:CIM_10Id ?labelCIM10.}
```

(1) et (2) espace de noms pour le modèle RDFS et le vocabulaire de l'ontologie ; (3) sélection de la variable à afficher en résultat sous la forme du compte des résultats correspondant à la requête ; (4) motif de la requête : nous cherchons dans les concepts ceux qui ont une annotation CIM\_10Id qui correspond à l'alignement du concept avec un code de la CIM-10. En résultat, nous affichons le nombre de concepts associés à un code CIM-10.

### **Extraire les concepts du DSM 5 qui ont un label DSM IV et CIM-10**

```
(1) PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
(2) PREFIX ont: <http://www.limics.fr/OntoPsychia_Classifications#>

(3) SELECT ?concept ?DSMIV ?CIM10
(4) WHERE {
(5) ?concept ont:DSM_IVId ?DSMIV.
(6) ?concept ont:CIM_10Id ?CIM10.}
```

(1) et (2) espace de noms pour le modèle RDFS et le vocabulaire de l'ontologie ; (3) sélection des variables à afficher en résultat ; (4)(5)(6) motif de la requête : nous cherchons dans les concepts ceux qui ont une annotation DSM IV et une annotation CIM-10. En résultat, nous affichons les concepts associés à leur code DSM IV et leur code CIM-10.

### 2.2.3 Les vocabulaires d'ontologies

#### Simple Knowledge Organization System (SKOS)

SKOS Core est un vocabulaire RDF qui permet de représenter la structure et le contenu d'une architecture conceptuelle, ici au sens d'un SOC, tels que le sont les thésaurus, classifications, terminologies, glossaires ou encore ontologies. Ce vocabulaire est publié et maintenu par le groupe de travail « Semantic Web Best Practices and Deployment » du W3C<sup>12</sup> [Miles *et al.*, 2005]. Au sein d'une ontologie, SKOS est utilisé pour gérer la hiérarchie conceptuelle (la ressource termino-ontologique, RTO). Il va ajouter des informations lexicales à chaque ressource conceptuelle, à chaque concept de l'ontologie à travers la notion de Label. En effet, la première caractéristique d'un concept est sa réalisation dans le langage naturel.

**Les Labels SKOS** associent à chaque concept ces réalisations linguistiques (un ou des signifiants) dans une langue défini par l'ontologue au travers des propriétés suivantes :

- `skos:prefLabel` : ce label existe une seule fois pour chaque langue de chaque concept.

```
skos:prefLabel "préoccupé"@fr
skos:prefLabel "preoccupied"@en
```

- `skos:altLabel` : ce label permet d'associer au concept des abréviations, acronymes, ou encore synonymes.

```
skos:prefLabel "inquiet"@fr
skos:prefLabel "inquiète"@fr
skos:prefLabel "préoccupée"@fr
skos:prefLabel "worried"@en
```

- `skos:hiddenLabel` : ce label est utilisé pour ajouter des formes lexicales à la ressource. Ces formes seront accessibles à l'application qui utilise l'ontologie, mais non visible autrement. Ce label permet par exemple d'ajouter des formes lexicales tronquées ou mal orthographiées.

```
skos:prefLabel "inquiète"@fr
skos:prefLabel "preoccupee"@fr
skos:prefLabel "worried"@en
```

**Les relations en SKOS** sont partagées en trois catégories, selon qu'elles définissent un lien hiérarchique de subsumption ou d'association non hiérarchique.

- `skos:broader` et `skos:narrower` : d'un concept plus général à un concept moins général et inversement. Cette relation est considérée comme transitive, mais cette transitivité n'est pas prise en charge par SKOS.

---

12. <https://www.w3.org/2001/sw/BestPractices/>

```
skos:prefLabel "maladie"@fr
skos:narrower ex:schizophrénie.
```

```
skos:prefLabel "schizophrénie"@fr
skos:broader ex:maladie.
```

- skos:related : entre deux concepts ayant un lien sémantique évident. Cette relation est symétrique.

```
skos:prefLabel "maladie"@fr
skos:related ex:médical.
```

```
skos:prefLabel "médical"@fr
skos:related ex:maladie.
```

**Les notes en SKOS** servent à définir les concepts de manière informelle, à l'intention des humains. Cela permet d'ajouter des définitions (ici avec la syntaxe XML) :

```
<skos:prefLabel xml:lang="en">commuting accident</skos:prefLabel>
<skos:prefLabel xml:lang="fr">accident de trajet</skos:prefLabel>
<skos:definition>Accident survenant pendant un trajet
domicile-travail/travail-domicile.</skos:definition>
```

mais également des exemples, des informations sur la partie éditoriale ou encore sur l'évolution du modèle. Compilé au Dublin Core, SKOS peut également décrire entièrement un SOC, son contenu et sa hiérarchie [Miles *et al.*, 2005]. Ces éléments sont décrits en totalité par la norme ISO 15836 :2009

### Le Dublin Core

Le Dublin Core est un modèle de description de ressources électroniques qui prend la forme d'un vocabulaire composé de 15 éléments, aussi appelés propriétés [The Dublin Core Metadata Initiative - DCMI, 2012]. Il est décrit par les normes ISO 15836 :2009<sup>13</sup> ; ANSI/NISO Standard Z39.85-2012<sup>14</sup> et IETF RFC 5013. Ce vocabulaire a été standardisé en 1998, par une suite de recommandations rédigées par un groupe de travail composé de spécialistes du monde de la librairie et de la communauté de chercheurs en librairie digitale [Weibel, 1997]. Aujourd'hui, c'est la Dublin Core Metadata Initiative qui maintient et supporte le développement du Dublin Core. Le Dublin Core a été développé avec plusieurs objectifs, dont celui de disposer d'un vocabulaire simple, utilisable par tous, y compris par des non-bibliothécaires. Ainsi, chaque personne est en mesure de décrire ses ressources, de manière à ce qu'elles soient visibles par des outils de recherches. Un autre objectif intéressant défendu par Weibel [1997] est l'interopérabilité sémantique. En effet, le Dublin Core permet d'unifier la description de ressources qui utilisent en complément d'autres formats descriptifs. Il est utilisé à la Bibliothèque nationale de France (BnF) en association avec d'autres formats bibliographiques « pour gérer et cataloguer les *signets* » (voir un extrait du code source à l'annexe F) « et pour accroître sur Internet la visibilité des catalogues et des collections numérisées. » [BnF, 2015].

13. <https://www.iso.org/fr/standard/52142.html>

14. <http://www.niso.org/standards/z39-85-2012>



## 2.3 Les composants de la modélisation ontologique

---

**Dublin Core non-qualifié, ou simple** est constitué des 15 éléments qui constituent le socle de base du vocabulaire [The Dublin Core Metadata Initiative - DCMI, 2012, BnF, 2015] et qui portent sur :

1. le contenu : titre (dc :title), sujet (dc :subject), description (dc :description), source (dc :source), langage (dc :language), relation (dc :relation : pour faire référence à une ressource apparentée. Exemple : un périodique dont est issu un article.), couverture (dc :coverage : « Périmètre ou domaine d'application du contenu de la ressource, c'est à dire ou la couverture spatiotemporelle de la ressource. » [BnF, 2015]) ;
2. la propriété intellectuelle : créateur (dc :creator), contributeur (dc :contributor), éditeur (dc :publisher), gestion des droits (dc :rights) ;
3. l'instanciation : date (dc :date), type (dc :type), format (dc :format), identifiant de la ressource (dc :identifier).

Chacun de ses éléments est associé à une recommandation de bonne pratique pour uniformiser leur utilisation. Par exemple, pour la date il est recommandé « d'encoder la valeur conformément au profil défini dans la norme ISO 8601 ou recommandation Date and Time Formats du W3C, qui comprend (notamment) des dates suivant la forme AAAA-MMJJ » [BnF, 2015] ; pour l'éditeur il est préférable de l'identifier par son ISBN.

**Dublin Core qualifié** est une extension du Dublin Core simple. Il inclut donc les éléments du Dublin Core simple et contient trois éléments en plus (dc :audience, dc :provenance, dc :rightsHolder) , des recommandations supplémentaires ainsi que des *raffinements*. Les raffinements servent à restreindre la signification d'un élément. Ce dernier conserve sa signification première, mais un sens plus spécifique vient s'y ajouter. Par exemple, l'élément *date* peut être *raffiné* par les éléments de raffinements suivants [Dublin Core Qualifiers, 2012] : Created (date de création de la ressource), Valid (date de validité de la ressource), Available (date de début de validité de la ressource), Issued (date de parution officielle de la ressource), Modified (date de modification de la ressource). Ces éléments de raffinement sont optionnels et peuvent être utilisés les uns indépendamment des autres.

## 2.3 Les composants de la modélisation ontologique

Une ontologie est construite à partir de trois composants : les classes qui sont les concepts organisés hiérarchiquement dans l'ontologie, les attributs ou propriétés qui décrivent les concepts par le biais de relations et les restrictions qui limitent l'interprétation des attributs [Noy et al., 2001]. Ces composants peuvent être écrits manuellement ou représentés dans une ontologie computationnelle en utilisant un langage d'ontologie, dont ceux présentés en section 2.2.2.

### 2.3.1 Notion de classes et d'instances de classe

Les classes correspondent aux nœuds de l'arbre taxonomique [Noy et al., 2001]. Elles sont la structure hiérarchique conceptuelle de l'ontologie. Une classe permet de regrouper des ressources qui ont des caractéristiques similaires. Les classes sont associées à une **intention**, une description sémantique sous la forme de restrictions ou propriétés, et à



une **extension**, l'ensemble des instances qui répondent à l'intention de la classe. Une classe est donc un concept du monde réel, tel que *Médecin*, *Maladie* ou encore *Patient* dont l'intention est : « personne affectée par une maladie ». Et l'extension est l'ensemble des personnes qui répondent à cette définition. Une ontologie peut ne contenir aucun individu, aucune instance, alors qu'elle contient obligatoirement des classes. L'intention d'une classe est donc décrite à travers une liste d'instances, et l'extension par une liste de restriction. Une classe peut être définie quand elle est dérivée d'autres classes ou primitive quand elle ne l'est pas. Par exemple, on peut créer une ontologie dans laquelle par relation de subsomption, la classe Humain est définie par l'union des classes primitives Femme et Homme :

$$\text{Humain} \subseteq \text{Femme} \cup \text{Homme}$$

La limite entre instance et classe peut sembler parfois un peu floue et se décide principalement selon le niveau de granularité souhaité dans l'ontologie. En effet, une grippe peut être instance de la classe maladie, mais peut également être une sous classe de la classe maladie, car la grippe se définit elle-même par un ensemble de différents virus. Nous insistons donc ici sur le fait que le langage de modélisation oblige lui-même à faire des choix préalablement au développement du modèle [Smith *et al.*, 2004].

### 2.3.2 Notion de propriétés, d'attributs, de rôles

**La relation de subsomption (*is a*)** caractérise la hiérarchie taxonomique de classes et sous-classes dans une ontologie. Un « Humain » est un « Être vivant ». La classe « humain » est une sous classe de la classe « Être vivant », donc la classe « Être vivant » subsume la classe « Humain ». Toutes les instances de « Humain » sont aussi instances de « Être vivant ».

**La relation de méronymie (*est une partie de*)** est une autre relation couramment modélisée dans une ontologie. Un « organe vital » est une partie du « corps humain ». Dans ce cas les instances de l'un ne sont pas les instances de l'autre, car les deux classes ne partagent pas les mêmes propriétés.

**Les relations binaires** permettent de lier les classes de deux manières : par une relation qui relie une instance de classe à une donnée (les *dataproperties*) ou par une relation qui relie deux instances de classe (les *objectproperties*). Pour cela nous définissons un domaine (sujet) et un co-domaine (objet) à la relation. Si le domaine ou le co-domaine sont défini à l'aide de plusieurs classes, ils deviennent l'intersection des classes.

Ces relations peuvent être définies tout comme les classes au travers de quantificateurs pour spécifier la relation entre les classes. Par exemple un patient est défini par le rôle de personne malade :

$$\begin{aligned} \text{Patient} &\subseteq \text{Humain} \\ &\cap \forall a \text{ PourMaladie.Maladie} \end{aligned}$$

## 2.4 Synthèse

---

Ce concepts indique que toutes les instances de Humain reliées par la relation aPourMaladie seront reliées à une instance de la classe Maladie. Cette relation n'est pas définie comme équivalente, car un humain malade n'est pas forcément un patient. Il doit être engagé dans un parcours de soin pour l'être. Nous pouvons ajouter :

$$\begin{aligned} Patient &\equiv Humain \\ &\cap (\forall aPourMaladie.Maladie) \\ &\cap (\forall aPourSuiviMedical.SuiviMedical) \end{aligned}$$

Ainsi, chaque instance de la classe Humain reliée par la relation aPourMaladie à une instance de la classe maladie et reliée par la relation aPourSuiviMedical à une instance de la classe SuiviMedical sera considéré par le raisonneur comme une instance de la classe Patient.

Les relations permettent de donner du sens au vocabulaire de l'ontologie, composé des classes et de leurs instances. Elles construisent l'interprétation du modèle du domaine.

## 2.4 Synthèse

Les ontologies sont de puissants outils de modélisation, qui viennent enrichir les SOC, dans le but de répondre à des problématiques d'organisation des connaissances.

Dans ce chapitre, nous avons défini l'ontologie en tant que représentation formelle d'un domaine du monde réel, des entités de ce domaine et des relations entre ces entités. Une ontologie permet de définir un vocabulaire commun et une représentation consensuelle d'un domaine donné. Les ontologies sont représentées par des graphes porteurs d'informations sémantiques. Elles permettent de partager de l'information aussi bien au niveau humain qu'au niveau machine. Les ontologies ont d'ailleurs été popularisées bien avant l'essor du Web Sémantique, grâce au besoin d'intégration de données de disciplines majeures, telles que la biologie et la médecine dans les années 1990. Aujourd'hui, leur rayonnement et les technologies qui y sont liées font partie intégrante du développement du Web Sémantique. Et des disciplines habituellement laissées à l'écart des nouvelles technologies, telle la psychiatrie, s'intéressent à leur puissance de modélisation, de partage de l'information et de raisonnement.

Nous verrons dans le chapitre suivant les techniques dont nous disposons pour créer, construire, développer des ontologies informatiques.



# Chapitre 3

## Construction d'ontologies informatiques

### Sommaire

<b>3.1 Engagements méthodologiques des ontologues et recommandations pour la construction d'ontologies</b>	<b>60</b>
3.1.1 Engagement sémantique, ontologique et computationnel	60
3.1.2 Recommandations à l'attention des ontologues	61
3.1.3 Processus de développement des ontologies	61
<b>3.2 Développement du modèle par approche ascendante (bottom-up)</b>	<b>62</b>
3.2.1 Méthodes « manuelles »	63
3.2.2 Méthodes basées sur l'acquisition de termes à partir de textes	66
3.2.3 Les logiciels pour l'acquisition automatique de termes à partir de textes	69
3.2.4 Comparaison des extracteurs de termes candidats YATEA et BIOTEX	72
<b>3.3 Développement du modèle par approche descendante (top-down)</b>	<b>73</b>
3.3.1 Le projet Common Kads	74
3.3.2 Le projet Sensus	75
<b>3.4 Développement du modèle par approche hybride (bottom-up et top-down)</b>	<b>77</b>
3.4.1 Macao	77
3.4.2 ToReuse2Onto	77
<b>3.5 Discussions sur les méthodes de construction d'ontologies</b>	<b>78</b>
<b>3.6 La modularité ontologique : l'ergonomie au service du développement du modèle.</b>	<b>79</b>
3.6.1 Définition de la modularité ontologique	79
3.6.2 Objectifs de la modularité ontologique	80
3.6.3 La composition modulaire	81
3.6.4 La décomposition modulaire	82
<b>3.7 Synthèse</b>	<b>83</b>

*Le chapitre précédent nous a permis d'approfondir l'objet principal de l'ingénierie des connaissances (IC) : l'ontologie et ses enjeux applicatifs dans le monde de l'intelligence artificielle (IA). Un autre pan de l'ingénierie des connaissances s'intéresse aux méthodes de développement de ces artefacts : choix de la modélisation sémantique, acquisition des connaissances ou encore choix des outils de construction. Les questions autour de ces méthodes restent encore vastes et soumises à des enjeux applicatifs : comment recueillir les connaissances à modéliser ? Est-il préférable d'utiliser des modèles déjà existants ? Si aucun modèle n'existe au préalable, comment développer l'ontologie ? Le développement du modèle ontologique peut être étudié selon deux approches à la fois opposées et complémentaires : ascendante et descendante. L'approche ascendante (bottom-up) se concentre sur la représentation abstraite de données brutes, alors que l'approche descendante (top-down) met l'accent sur la réutilisation de modèles déjà existants. Dans ce chapitre sur la construction d'ontologies, nous nous arrêtons dans une première section sur la notion « d'engagement sémantique », complétée par des recommandations pour la construction des ontologies. Dans la deuxième section, nous explorons des méthodologies par approche ascendante. Dans la troisième section, nous explorons celles par approche descendante. Et dans la quatrième section, nous explorons les méthodologies par approche hybride. Enfin, dans la dernière section, nous faisons un point sur l'apport de la modularité, pour un développement tenant compte des connaissances toujours croissantes.*

### 3.1 Engagements méthodologiques des ontologues et recommandations pour la construction d'ontologies

#### 3.1.1 Engagement sémantique, ontologique et computationnel

**Bachimont [2000]**, dans le cadre du développement de la méthode ARCHONTE (présentée en section 3.2.2) propose une définition originale de la modélisation d'ontologies, caractérisée par trois niveaux d'engagement : sémantique, ontologique et computationnel.

**L'engagement sémantique** consiste à définir un certain nombre de primitives de représentation, propres au domaine que nous souhaitons modéliser. Ces primitives sont alors les concepts de l'ontologie, liés à un libellé linguistique de la langue du domaine. Les concepts sont discriminés par le principe différentiel, qui permet de déterminer la signification d'un concept selon sa position dans l'arbre (relation de subsumption), par identités et différences avec ses voisins (concept *parents* et concept(s) *frère(s)*). L'engagement sémantique est résumé comme tel par l'auteur : « ensemble de prescriptions interprétatives qu'il faut respecter pour que le libellé fonctionne comme une primitive ». Cette démarche de différenciation des concepts entre eux peut s'apparenter à une analyse sémique telle que décrite en section 1.2.

**L'engagement ontologique** définit l'extension des concepts, soit les objets qui répondent à la définition sémantique du concept. Il sert à modéliser les instances des concepts, ou les nouveaux concepts qui vont pouvoir être modélisés par intersection de la liste des instances partagées par deux mêmes concepts. L'auteur prend l'exemple suivant : si dans une

### 3.1 Engagements méthodologiques des ontologies et recommandations pour la construction d'ontologies

---

ontologie nous avons un concept « Acteur » et un concept « Être humain » toutes les instances qui sont à l'intersection de ces concepts sont des « Personne-Acteur ». Ce nouveau concept ne devient pas un concept sémantique répondant au principe de l'engagement sémantique, mais est un concept formel défini par son extension et ses concepts parents.

$$\begin{aligned} \text{PersonneActeur} &\equiv \text{Personne} \\ &\cap (\forall a \text{ PourMetier. Acteur}) \end{aligned}$$

Ces concepts formels existent grâce à l'engagement ontologique. L'ontologie résultante de cet engagement n'a plus la forme d'un arbre, mais la forme d'un treillis.

**L'engagement computationnel** correspond au niveau axiomatique, c'est à dire au niveau des opérations réalisables sur les concepts. Ces opérations confèrent aux concepts leur sémantique d'un point de vue computationnel, calculatoire. Par exemple : [Etre humain : John Wayne] -> (a\_pour\_fonction) -> [acteur].

#### 3.1.2 Recommandations à l'attention des ontologues

**Aimé et Charlet [2013]** analysent les « points critiques » qui peuvent être rencontrés au début du développement collaboratif d'une ontologie. Les auteurs s'intéressent aux ontologies en tant (1) qu'« objet de la psychologie cognitive », car elles sont la représentation des connaissances mentales consensuelles propres aux individus et en tant (2) « qu'objet de la psychologie sociale », car elles permettent à partir des connaissances partagées par un groupe d'individus, d'établir un consensus autour de la compréhension d'un domaine. Ils se penchent sur les notions de normalisation et de conformisme qui entrent en œuvre au sein des ontologies. Ils rappellent à cette occasion, l'influence des experts sur le groupe et les individus qui le constitue. L'étude permet alors de poser un certain nombre de recommandations qui viennent compléter les engagements de **[Bachimont, 2000]** et étayer l'ontologue dans la construction des ontologies.

1. Privilégier l'approche collaborative : chaque personne ayant quelque chose de particulier à apporter au modèle.
2. Avoir un coordinateur-moderateur : pour assurer la gestion du groupe et garantir une modélisation adéquate.
3. Privilégier une approche ontologique de type modulaire : car chaque personne à quelque chose de particulier à apporter selon son domaine d'expertise.
4. Analyser l'écosystème : afin de cerner le fonctionnement du groupe et la représentation conceptuelle du domaine propre à chacun.
5. Choisir les experts : une personne reconnue dans son domaine de compétence, voire plusieurs personnes allant du niveau junior au niveau senior afin de faire varier le niveau de représentation.

#### 3.1.3 Processus de développement des ontologies

La construction d'ontologies met en jeu différentes étapes qui permettent de transformer des données en modèles ontologiques, qui seront ensuite intégrés dans un système à base de connaissances (SBC).

1. **Évaluation des besoins** : formulation des besoins liés à l'ontologie et définition de la granularité (que nous avons présenté en introduction de ce manuscrit) : [Noy *et al.*, 2001] rappelle très justement l'importance de répondre aux questions suivantes avant d'entamer le processus de modélisation et le développement de l'ontologie : « Quel domaine va couvrir l'ontologie ? » ; « Dans quel(s) but(s) utiliserons nous l'ontologie ? » ; « À quels types de questions l'ontologie devra-t-elle fournir des réponses ? » ; « Qui va utiliser et maintenir l'ontologie ? ». Ces questions permettent de limiter la portée du modèle, la granularité ou encore d'orienter la conceptualisation.
2. **Recueil des connaissances** : qui constitueront la base de connaissances à modéliser. Cette étape est certainement la plus ardue en ingénierie des connaissances, étant donné l'importance des facteurs extérieurs qui conditionnent sa réussite. Tel que le rappel [Schvartz *et al.*, 2007] « la disponibilité de l'expert est cruciale, mais pas toujours assurée ». Au prétexte de l'ingénierie des connaissances, l'acquisition des connaissances était vu comme un transfert de données de la tête d'un expert à un système organisé. Force est de constater avec les difficultés que représente l'acquisition des connaissances, qu'il en est tout autre. C'est aussi pour cette raison que de plus en plus de systèmes sont développés pour réaliser cette tâche de manière automatique et indépendante de tout expert [Schreiber *et al.*, 1994, Schvartz *et al.*, 2007]. Ce point est abordé dans les *chapitres 5 et 6* de ce manuscrit.
3. **Développement du modèle** : avec « construction d'un schéma de modèle conceptuel » et « définition du modèle conceptuel » [Charlet et Bachimont, 1998] : par approche ascendante, descendante ou hybride. Le développement de notre ontologie est décrite au *chapitre 5* et au *chapitre 6*.
4. **Validation du modèle** : par « implémentation de ce dernier dans une base de connaissances opérationnelle » [Charlet et Bachimont, 1998]. La validation est également réalisée selon les techniques et méthodes présentées au *chapitre 4*. Dans le cadre de notre projet, nous avons développé notre propre méthode de validation d'ontologies, présentée et expérimentée au *chapitre 7*.

### 3.2 Développement du modèle par approche ascendante (bottom-up)

Les méthodes ascendantes se concentrent sur la définition et l'identification de besoins initiaux, qui guident ensuite l'analyse de données et le développement du modèle conceptuel. Ces méthodes mettent en œuvre des techniques pour le recueil de données, l'extraction d'informations, la fouille de connaissances, ou encore la structuration des données [Charlet et Bachimont, 1998]. Le point central de ces méthodes est l'acquisition des connaissances du domaine, qui seront modélisées dans l'ontologie. Actuellement deux méthodes dominent : l'une s'appuie sur des terminologies existantes, et l'autre sur des outils du Traitement Automatique des Langues (TAL). La figure 3.1 illustre ce cheminement qui va du recueil des connaissances expertes aux modèles et structures en réseau.

### 3.2 Développement du modèle par approche ascendante (bottom-up)

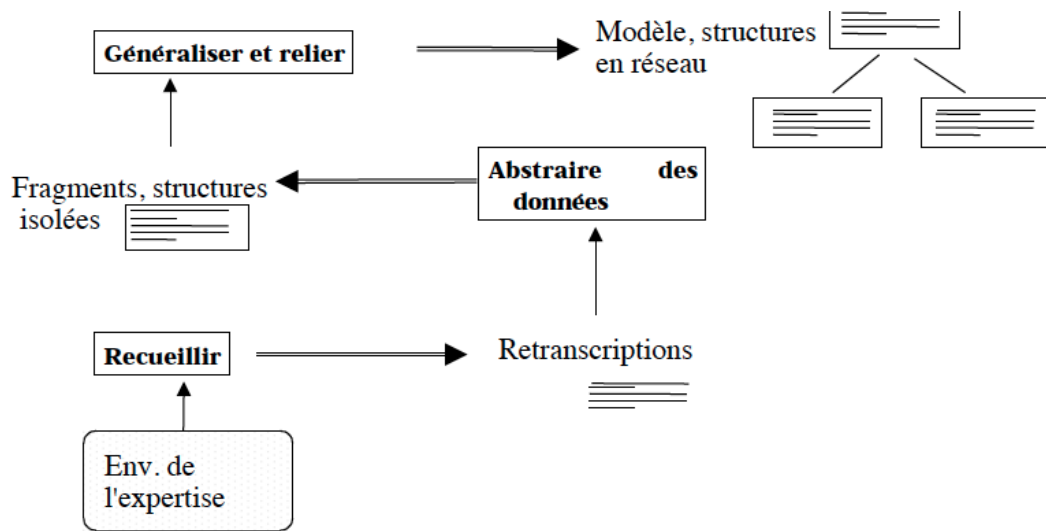


FIGURE 3.1 – Des données vers un modèle structuré Aussenac-Gilles et Charlet [2010].

#### 3.2.1 Méthodes « manuelles »

Les premières méthodes d'aide à la construction d'ontologies qui ont vu le jour à partir des années 1990 étaient dites « manuelles ». En opposition aux méthodes qui utilisent des traitements automatiques pour construire tout ou partie de la base de concepts. Elles proposent un processus très détaillé pour guider l'ontologue durant le développement de son ontologie, sans assistance logiciel autre que les éditeurs d'ontologies.

##### La méthodologie de Unshold et King

En 1995, à l'Artificial Intelligence Applications Institute (AIAI) de l'université d'Edinburgh est développée une ontologie pour modéliser l'entreprise. King et Uschold [1995] font alors état d'un manque de consensus et de clarté autour des questions de méthodologie pour la construction d'ontologies. En se basant sur la littérature, les chercheurs proposent donc un socle méthodologique (illustré en figure 3.2) composé de quatre étapes, pour aider à la construction d'ontologies. Fernández-López et Gómez-Pérez [2002] précisent qu'ils sont ainsi les premiers à pointer l'intérêt d'utiliser des méthodologies pour le développement d'ontologies.

1. Identifier les objectifs : de manière à savoir pourquoi l'ontologie est développée et dans quel(s) but(s) elle sera utilisée.
2. Construire l'ontologie :
  - Capturer les connaissances : en identifiant les concepts et les relations du domaine à modéliser, ainsi que les termes associés.
  - Coder l'ontologie : en modélisant les connaissances définies à l'étape précédente, dans un langage de formalisme d'ontologies.
  - Intégrer des ontologies existantes : aussi bien à l'étape de capture des connaissances qu'à l'étape de modélisation des connaissances.



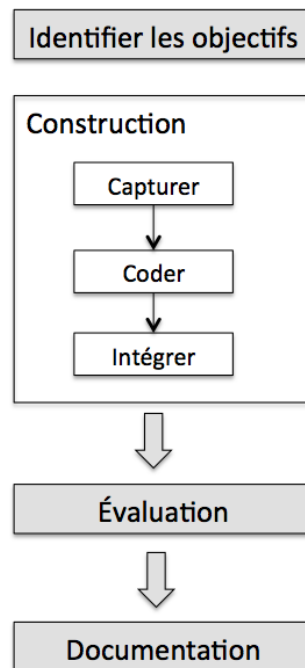


FIGURE 3.2 – Méthodologie de King et Uschold [1995] d'après Fernández-López et Gómez-Pérez [2002].

3. Évaluer : selon la définition de Gómez-Pérez *et al.* [1995] : « make a technical judgement of the ontologies, their associated software environment, and documentation with respect to a frame of reference. The frame of reference may be requirements specications, competency questions, and/or the real world. » – « Faire un jugement technique des ontologies, de leur environnement de développement et de la documentation en respectant un cadre de référence. Le cadre de référence peut être les spécifications des exigences, les questions de compétence et/ou le monde réel ».
4. Documenter : le contenu de l'ontologie.

Les auteurs ajoutent que ces étapes doivent également définir un ensemble de techniques, méthodes, principes et lignes directives, ainsi que des liens explicites entre elles.

### Methontology

METHONTOLOGY a été développée dans le laboratoire d'Intelligence Artificielle de l'Université Polytechnique de Madrid. Fernández-López *et al.* [1997] proposent une aide méthodologique en dix étapes, pour guider l'ontologue durant le développement manuel de l'ontologie. La méthode est supportée par l'éditeur mono-utilisateur ODE et l'éditeur multi-utilisateurs WEBODE. Le but est d'offrir un cadre pour permettre aux utilisateurs de développer une ontologie complète et correcte [Fernández-López et Gómez-Pérez, 2002]. La méthode détaillée ci dessous est illustrée figure 3.3.

1. Les activités de gestion de projet :
  - L'ordonnancement : liste les tâches à effectuer, avec leur agencement, leur temps de développement et les ressources qui seront nécessaires à leur élaboration.

### 3.2 Développement du modèle par approche ascendante (bottom-up)

---

- Le contrôle : est effectué sûr chaque tâche pour suivre le bon déroulement de l'ordonnancement.
  - La qualité : valide que chaque produit développé satisfait ses exigences.
2. Les activités axées sur le développement :
- La spécification : définit les buts de l'ontologie, les utilisations attendues et les utilisateurs finaux.
  - La conceptualisation : est réalisée en plusieurs étapes, elle définit la structure des connaissances du domaine dans le modèle [Fernández-López *et al.*, 1999] :
    - (a) Construire le glossaire des termes qui seront inclus dans l'ontologie, préciser leur définition en langage naturel, identifier leurs synonymes et leurs acronymes
    - (b) Construire des taxinomies de concepts
    - (c) Construire des diagrammes de relations binaires
    - (d) Construire le dictionnaire de concepts qui inclut, pour chaque concept, ses propriétés d'instance, ses propriétés de classe et ses relations
    - (e) Décrire en détail chaque relation binaire
    - (f) Décrire en détail chaque propriété d'instance
    - (g) Décrire en détail chaque propriété de classe
    - (h) Décrire en détail chaque constante (les constantes donnent des informations sur le domaine de connaissances)
    - (i) Décrire les axiomes formels
    - (j) Décrire les règles utilisées pour contraindre le contrôle et pour inférer des valeurs aux propriétés.
  - La formalisation : traduit le modèle conceptuel dans un modèle formel.
  - L'implémentation : implémente le modèle formel en modèle computationnel.
  - La maintenance : permet de mettre à jour et corriger l'ontologie le cas échéant.
3. Les activités de support : sont réalisées en parallèle aux activités axées sur le développement.
- L'acquisition de connaissances : vise à acquérir les connaissances du domaine à modéliser. Cette étape est également réalisée manuellement, par le biais d'analyses textuelles manuelles et de rencontres avec des experts.
  - L'évaluation : permet de juger de l'adéquation de l'ontologie au fur et à mesure de son développement.
  - L'intégration de concepts : permet l'ajout d'autres ontologies.
  - La documentation : détaille chaque étape du développement de l'ontologie.
  - La gestion de la configuration : enregistre toutes les versions de l'ontologie pour en contrôler les changements et évolutions.

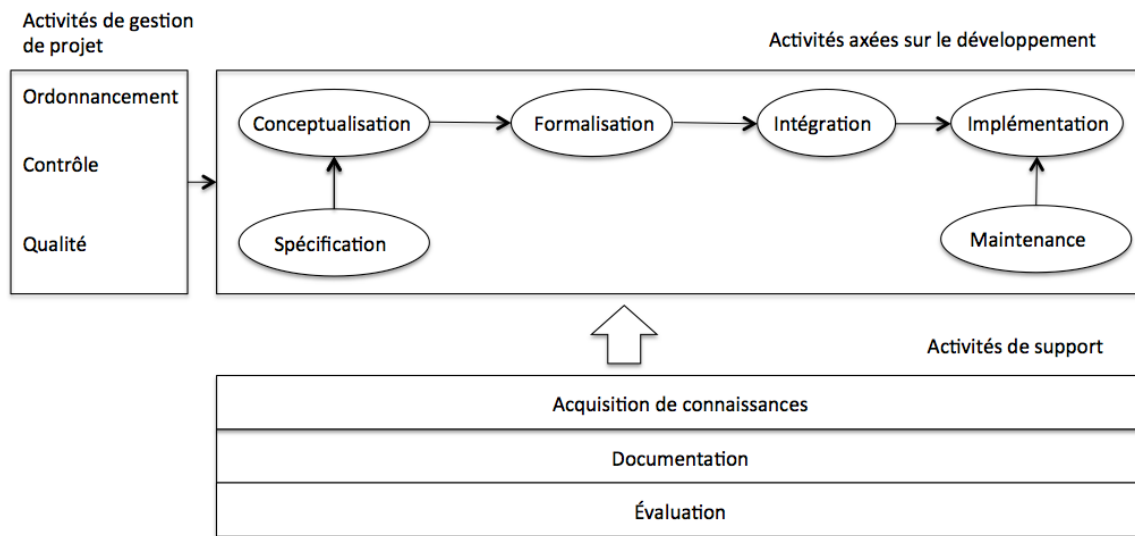


FIGURE 3.3 – Methontology

### 3.2.2 Méthodes basées sur l'acquisition de termes à partir de textes

Les méthodes basées sur l'acquisition de termes à partir de textes utilisent des sources de connaissances construites en amont par un logiciel (voir en section 3.2.3), puis validées par des humains. Ces méthodes sont dites semi-automatiques, car la base de termes relative au domaine à modéliser est construite grâce à une machine, avant l'intervention humaine qui valide ou non les termes en concepts.

**La méthode ARCHONTE (ARCHitecture for ONTological Elaborating) de [Bachimont *et al.*, 2002, Charlet *et al.*, 2006]**

Cette méthode s'articule en trois étapes qui correspondent aux trois engagements méthodologiques présentés en 3.1.1. Cette méthode est illustrée dans la figure 3.4 : après la construction du corpus, la phase de normalisation permet de développer une ontologie « différentielle », ensuite la phase de formalisation permet de développer une ontologie « référentielle » et enfin la phase d'utilisation du formalisme permet de développer une ontologie « computationnelle ».

Charlet *et al.* [2006] reprend les grandes étapes de ARCHONTE pour détailler une méthode de conceptualisation en quatre étapes :

**Étape 1 : la construction d'un corpus** en lien avec le domaine à modéliser dans l'ontologie, permet d'étudier la terminologie de ce domaine, à l'aide d'outils d'analyse automatique. Ce point est détaillé plus avant lors de la présentation de notre corpus, à la section 5.2.2.

**Étape 2 : la normalisation sémantique** (engagement sémantique) s'inspire des réflexions et des travaux de Pottier [2001] et de Rastier *et al.* [1994]. Ce dernier a approfondi les notions de l'analyse sémique au sein de sa sémantique interprétative. La normalisation sémantique se sert de cette analyse pour désambiguïser les concepts, contraindre leur

### 3.2 Développement du modèle par approche ascendante (bottom-up)

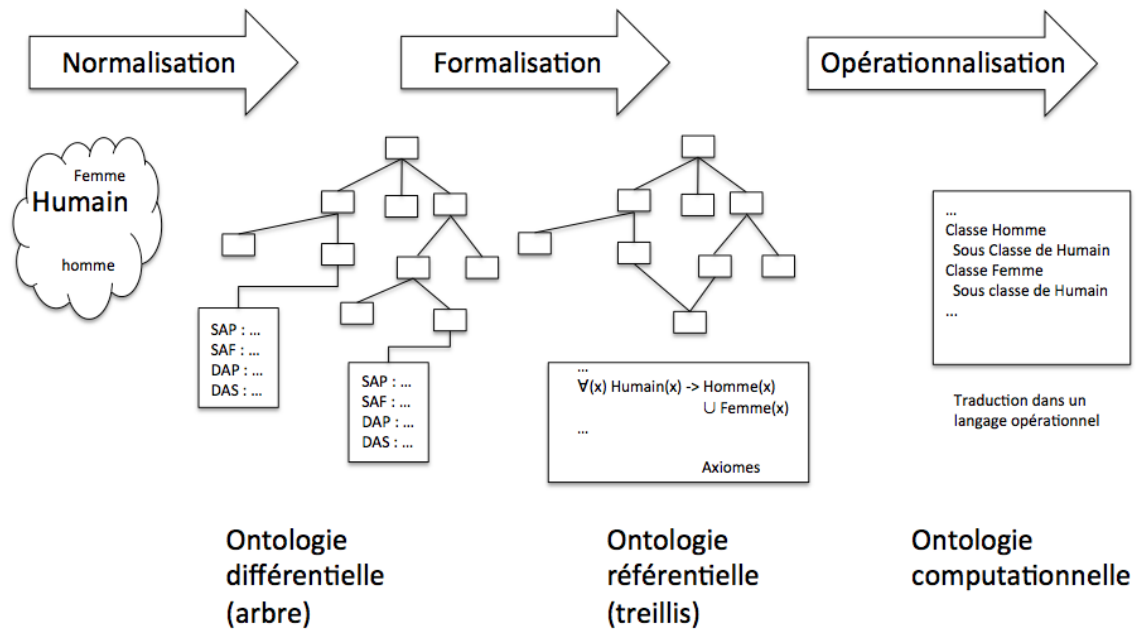


FIGURE 3.4 – La méthode ARCHONTE [Bachimont *et al.*, 2002].

interprétation dans l'ontologie. Le sens des concepts est déterminé par opposition et/ou identité à ses voisins, de la même manière que l'analyse sémique présentée en section 1.2.4 différencie des concepts. Ce travail de normalisation sémantique permet ainsi de fixer le sens du concept dans l'ontologie. Les termes *schizophrénie de type paranoïde* et *schizophrénie de type désorganisé* seront deux concepts distincts dans l'ontologie partageant le même concept parent *schizophrénie*. Cette normalisation sémantique permet aussi d'annuler l'effet de contexte. Par exemple, dans une ontologie de la nourriture, un « avocat » fera toujours référence au fruit. Cependant, dans une ontologie sur le droit, il fera référence à une profession et/ou une personne humaine. En outre, ce principe différentiel impose la représentation en arbre des ontologies. Car, si un concept se définit par rapport à ses voisins, il hérite des sèmes d'un concept parent, mais ne dispose pas des sèmes de ses concepts frères et dispose au moins d'un sème en moins que ces concepts enfants. Cela revient également à dire qu'un concept ne peut avoir qu'un parent. Car si un concept à deux parents, cela signifie qu'il hérite de sèmes contradictoires ou de sèmes complémentaires qui ne justifierait pas la création de deux concepts. La normalisation permet de passer d'une liste de termes à une ontologie différentielle.

**Étape 3 : la formalisation des connaissances** (engagement ontologique) consiste à formaliser des connaissances en définissant les domaines (sujets) et co-domaines (objets) des relations ou en définissant l'extension des concepts (les instances) qui peuvent entraîner la création de nouveaux concepts. Cette étape permet de passer d'une ontologie différentielle (constituée d'un ensemble de termes normalisés en concept) à une ontologie référentielle (constituée des instances de concepts).

**Étape 4 : l'utilisation de la formalisation** (engagement computationnel) sert à opérationnaliser les connaissances dans un langage d'ontologies et de règles logiques (plus de détails en section 2.2.2). L'ontologie est alors computationnelle et fondée sur une sémantique dite « opérationnelle ».

La méthode ARCHONTE a notamment été utilisée dans la cadre du développement des ontologies ONTO-PNEUMO [Baneyx, 2007] ou ONTO-URGENCE [Charlet *et al.*, 2012]. Pour le développement d'une ontologie de domaine de la médecine périnatale, Dhombres *et al.* [2010] ont enrichie la méthode ARCHONTE d'une étape descendante, pour la réutilisation de ressources existantes.

### OntoLearn Reloaded

La méthodologie ONTOLEARN RELOADED de Velardi *et al.* [2005, 2013] vise au développement automatique de taxonomies à partir de corpus et de sites Web. Le processus se décompose en cinq étapes :

1. Extraction de la terminologie d'un corpus de domaine, via un extracteur de termes, le processus est détaillé dans [Navigli et Velardi, 2004]) ;
2. Extraction de définitions et hyperonymes associés aux concepts au travers de (1) taxonomies, (2) documents disponibles sur le web et (3) outils d'analyse de phrases définitoires. Pour chaque terme candidat, un classificateur indépendant du domaine est utilisé pour sélectionner des définitions liées avec « is a » (e.g. Un avocat est un fruit) à partir des phrases candidates et extraire les hyperonymes<sup>1</sup> correspondants du terme après analyse de la phrase. Le processus est détaillé dans [Velardi *et al.*, 2013].
3. Filtrage des définitions qui ne correspondent pas au domaine de la taxonomie. Puis construction du graphe des relations de subsomption à l'aide des liens d'hyperonymies.
4. Élagage du graphe, afin de ramifier les hyperonymes multiples et détacher les nœuds mal étiquetés ou qui ne correspondraient pas au domaine.
5. Recouvrement d'une partie de l'élagage réalisé à l'étape précédente.

### Text2Onto

La méthodologie TEXT2ONTO [Cimiano et Völker, 2005] propose de construire automatiquement une ontologie à partir de l'analyse d'un corpus en anglais, espagnol ou allemand. L'outil identifie les classes, les relations de subsomption, de méréologie, les instances, les propriétés d'objet avec domaine et co-domaine et les équivalences logiques. Pour faire ce travail, l'outil utilise des algorithmes différents pour chacun de ses artefacts ontologiques. Par exemple, pour construire les relations de subsomption l'outil utilise, tout comme ONTOLEARN RELOADED, la base lexicale WordNet. TEXT2ONTO se base aussi sur l'outil de traitement et d'analyse du langage GATE et la plateforme d'environnement

1. L'hyperonyme est sélectionné dans les synset de plus haut niveau de la base lexicale WordNet.

d'ingénierie ontologique NEON TOOLKIT<sup>2</sup> [Haase *et al.*, 2008]. Une fois les artefacts extraits, ils peuvent être implémentés en OWL, RDFS ou F-Logic.

#### TERMINAE

TERMINAE [Biebow *et al.*, 1999] est une plateforme d'aide à la construction de ressources termino-ontologiques (RTO) à partir de ressources textuelles, en français et en anglais. La dernière version, disponible en application Eclipse, a été développée dans le cadre de la participation du Laboratoire d'Informatique de Paris Nord (LIPN) à un projet européen, puis améliorée à l'occasion du projet ANR DAFOE [Szulman *et al.*, 2009, 2010]. La plateforme permet de développer une RTO en intervenant sur trois niveaux :

1. **Le niveau linguistique et terminologique** : le premier niveau permet de filtrer les termes candidats extraits par YATEA, SYNTAX-UPERY, ANNIE (disponible sous la plateforme GATE), ou encore TERMOSTAT (voir en section 3.2.3 pour plus d'informations concernant ces outils). À ce niveau, nous pouvons valider les termes, les enlever ou en ajouter. Ensuite, un niveau terminologique gère un ensemble d'informations linguistiques associées à chaque terme sélectionné : appelé « fiche terminologique ». Cette fiche contient les variantes du terme et ses occurrences dans les textes.
2. **Le niveau termino-conceptuel / la normalisation** : permet de développer une RTO exportable en Simple Knowledge Organisation System (SKOS) (voir en section 2.2.3) à partir de la terminologie construite à l'étape précédente. L'ontologue opère donc une normalisation sur ces données, qui va permettre d'obtenir un réseau de termes non ambigus et interconnectés par des relations taxonomiques et sémantiques [Omrane *et al.*, 2012]. Cette étape est réalisée manuellement.
3. **Le niveau conceptuel ou ontologique / la formalisation** : transforme le réseau terminologique en ontologie et permet ainsi de lier la terminologie du domaine à son modèle conceptuel [Omrane *et al.*, 2012]. Dans sa version actuelle, la plateforme TERMINAE intègre l'éditeur NEON TOOLKIT [Haase *et al.*, 2008]. Cependant, une fois les concepts définis, il est possible d'exporter le projet en OWL et de travailler sous un autre éditeur d'ontologies.

TERMINAE différencie ainsi trois niveaux de développement, qui vont de l'extraction des termes à la formalisation du modèle. Il ne s'agit pas d'un simple outil, mais bien d'une méthodologie qui reprend pas à pas les étapes propres au développement d'une ontologie.

#### 3.2.3 Les logiciels pour l'acquisition automatique de termes à partir de textes

##### Acquisition de termes via des thésaurus et les outils de balisage et d'extraction associés

Le méta-thésaurus<sup>3</sup> Unified Medical Language System (UMLS) [Bodenreider, 2004] fournit de nombreuses ressources pour l'extraction de termes médicaux selon des do-

---

2. [http://neon-toolkit.org/wiki/Main\\_Page.html](http://neon-toolkit.org/wiki/Main_Page.html)

3. À noter qu'un méta-thésaurus est composé de thésaurus qui ont été regroupés ensemble via un niveau supérieur de hiérarchisation.

maines spécialisés. Actuellement, 145 thésaurus<sup>4</sup> sont disponibles dans 20 langues dont l'anglais, le français, l'allemand, le japonais et le russe. En France, différents projets ont été développés tel que l'Unified Medical Lexicon for French (UMLF) [Zweigenbaum *et al.*, 2005], puis InterSTIS [Cartoni et Zweigenbaum, 2010], ou encore le Catalogue et Index des Sites Médicaux de langue Française (CISMeF) [Darmoni et Joubert, 2000]. Ils visent à définir, pour le français, des ressources aussi complètes que celles contenues dans l'UMLS. Ces méthodes s'appuient notamment sur des outils de repérage qui permettent de retrouver dans les documents, les termes contenus dans les thésaurus. Ces méthodes sont donc pertinentes lorsqu'on a déjà une idée des concepts que l'on veut extraire de nos documents, et que l'indépendance face à des bases de connaissances pré-existantes n'est pas recherchée.

### Acquisition de termes via des extracteurs de termes candidats (ETC)

Cette approche s'appuie sur l'extraction de termes candidats en corpus spécialisés. Les outils développés utilisent principalement des techniques du traitement automatique du langage (TAL), telles que les analyses syntaxiques (permettant de reconnaître les phrases correspondant à la syntaxe d'une langue), les annotations morpho-syntaxiques (attribuant à chaque mot d'un texte son étiquette grammaticale), et les méthodes de statistiques linguistiques, afin d'obtenir une liste de termes candidats à valider manuellement. Parmi les outils utilisant ces méthodes, nous pouvons citer BIOMedical Term EXtraction (BioTex) [Lossio-Ventura *et al.*, 2014], Yet Another Term extrActor (Yatea) [Aubin et Hamon, 2006, Hamon, 2012], SYNTAX-UPERY [Bourigault et Lame, 2002], TTC TERM-SUITE [Rocheteau et Daille, 2011] – anciennement ACABIT – ou THERMOSTAT [Drouin, 2003]. Ces méthodes sont à privilégier quand nous n'avons pas d'idée préalable des concepts que nous allons extraire, et que cette recherche ne doit pas être influencée par des bases de connaissances pré-existantes.

BioTEX a été développé dans le cadre de la thèse de Juan Antonio Lossio Ventura, dans le Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM)<sup>5</sup>. Cet outil s'appuie sur la combinaison de méthodes linguistiques et statistiques à l'aide de patrons syntaxiques et de différentes mesures statistiques (LIDF-value, L-value, C-value, Okapi BM25, TFIDF). Ce logiciel est disponible aussi bien pour le français que pour l'anglais et l'espagnol. L'extraction se fait en plusieurs étapes :

1. **Annotation morpho-syntaxique du corpus** : l'annotation morpho-syntaxique d'un texte consiste à attribuer à chaque mot une étiquette indiquant sa catégorie grammaticale (par exemple, verbe, nom, pronom).
2. **Extraction des termes candidats** : avant l'application de mesures statistiques, BioTex sélectionne les termes correspondants à des patrons linguistiques spécifiques. Ces derniers ont été établis à l'aide de l'observation des plus fréquentes structures syntaxiques des termes biomédicaux issus de l'UMLS (pour l'anglais) et du MeSH (pour le français) [Lossio-Ventura *et al.*, 2013].

4. <http://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>

5. <http://tubo.lirmm.fr/biotex/>



3. **Classement des termes candidats** : selon l'ordre d'adéquation au domaine à l'aide des mesures F-TFIDF-C ou F-OCapi (décrites dans [Lossio-Ventura \*et al.\* \[2014\]](#)) et de LIDF-value (décrite dans [Lossio-Ventura \*et al.\* \[2014\]](#)). Ces mesures permettent notamment de combiner des valeurs telles que la fréquence inverse de document (idf) qui permet de mesurer la rareté d'un terme, la distribution des mots dans le corpus, les informations linguistiques et statistiques, entre autres.
4. **Calcule des cooccurrences** : pour améliorer le classement des termes candidats (plus un mot a de voisins, moins il est considéré comme spécifique, car utilisé dans des contextes généraux). L'originalité de cet outil est donc la combinaison de différentes mesures statistiques pour classer les termes extraits avec les patrons syntaxiques. Le but est de proposer en premier à l'utilisateur les termes les plus pertinents de son corpus (par exemple, dans le cadre d'un corpus de psychiatrie, le candidat terme unigramme « émoussement » n'a pas d'intérêt, cependant le candidat terme 3-grammes « émoussement des affectes » est lui pertinent). Une aide à la validation est ensuite proposée en explicitant les termes candidats extraits du corpus qui sont aussi présents dans l'UMLS ou le MeSH-fr.

SYNTEX-UPERY<sup>6</sup> se compose de deux modules. L'analyseur syntaxique SYNTEX crée un réseau de dépendances entre les mots et les syntagmes. Chaque syntagme constitue un candidat terme et est caractérisé par une fréquence d'apparition dans le corpus. UPERY va ensuite permettre de rapprocher les termes du réseau et leurs contextes syntaxiques via des mesures de proximité distributionnelle.

TERMOSTAT<sup>7</sup> a la particularité de s'appuyer sur la mise en opposition d'un corpus spécialisé et non spécialisé. Il s'utilise uniquement en ligne et ne garantit donc pas la confidentialité des données.

TTC TERMESUITE<sup>8</sup> utilise des corpus bilingues comparables afin d'aligner les termes spécialisés.

YATEA<sup>9</sup> a été développé dans le cadre du projet ALVIS<sup>10</sup>. Il permet d'identifier des groupes nominaux qui peuvent correspondre à des termes spécialisés d'un corpus. Il fournit une analyse syntaxique dans un fichier xml, sous forme d'une décomposition en tête et modifieur. L'extraction des termes est réalisée avec des patrons d'analyse simple. Une désambiguïsation endogène est réalisée au préalable, puis des mesures de pondération statistique permettent de discriminer les termes candidats. YATEA prend en entrée des données étiquetées morphologiquement via l'annotateur TREETAGGER<sup>11</sup> pour le français. TREETAGGER est un étiqueteur morphosyntaxique développé par Helmut Schmid à l'Université de Stuttgart<sup>12</sup> en Allemagne.

---

6. <http://w3.erss.univ-tlse2.fr:8080/index.jsp?perso=bourigault&subURL=syntex.html>

7. <http://termostat.ling.umontreal.ca/>

8. <http://ttc.syllabs.com/>

9. <http://search.cpan.org/~thhamon/Lingua-YaTeA-0.5/>

10. <http://lipn.univ-paris13.fr/fr/rcln-projets/rcln/projets/alvis>

11. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>

12. <http://www.uni-stuttgart.de/home/>



### 3.2.4 Comparaison des extracteurs de termes candidats YATeA et BIOTeX

Au cours de notre étude, nous avons été amenés à utiliser ces deux extracteurs de termes. YATeA dans un premier temps, pour le développement du module des « facteurs sociaux et environnementaux », nous a semblé être la solution adéquate, car nous souhaitions travailler avec la plateforme TERMINAE. Cependant, cela n'a pas été possible à cause de la taille de notre corpus, qui outrepassait les limites de TERMINAE, ainsi que de YATeA. Nous nous sommes donc tournés vers l'ETC BIOTeX pour le développement du module des « maladies ».

#### Méthode

Afin de comparer ces deux logiciels de manière optimale, nous les avons étudiés sur le même corpus et dans le but d'extraire des termes candidats du même domaine. Le corpus utilisé est un extrait de notre corpus, composé uniquement de 5 003 mots. Ceci afin de rendre réalisable les divers traitements manuels par une seule personne. Nous avons commencé par réaliser une extraction manuelle des termes relatifs au domaine du médical (trouble, maladie, symptôme et traitement). Nous avons retenu 528 termes (sur 5 003 que compte le corpus test) pouvant faire partie d'une RTO. Nous avons ensuite réalisé l'extraction avec les deux ETC, en précisant pour BIOTeX que nous ne souhaitions pas extraire d'unigram, car YATeA ne propose pas l'extraction d'unigram. Ensuite, nous avons compté dans la liste des termes candidats proposés par chacun des extracteurs, le nombre de termes appartenant à la liste des termes validés manuellement. À ce compte, nous avons ajouté les termes candidats dérivés d'un terme validé manuellement. Par exemple, si nous avions validé manuellement : « altération mnésique antérograde » et que l'extracteur nous proposait le terme « altération mnésique » nous le comptons en terme candidat valide obtenu par dérivation.

#### Résultats

Pour calculer la performance technique des ETC, nous les avons testés sur notre corpus comparatif, ainsi que sur la totalité du corpus. Nous constatons la difficulté de traitement d'un gros corpus (les corpus de plusieurs Méga-octets et millions de mots). YATeA est en échec, et il faut plusieurs heures à BIOTeX pour arriver au bout de l'analyse. Cependant, les options proposées par BIOTeX à l'attention des gros corpus permettent de réduire ce temps.

TABLEAU 3.1 – Calcul de la performance des ETC.

Taille du corpus	TRETAGGER et YATeA	BIOTeX
33 ko	< 0.9 et = 3 scd	3 scd
30,5 mo	= 43 scd et échec à 10 min de traitement	18 heures

#### Définition des mesures utilisées dans le tableau 3.2 :

- Vrai positif : terme candidat valide et extrait par l'ETC.
- Faux positif : terme candidat non valide et extrait par l'ETC.

### 3.3 Développement du modèle par approche descendante (top-down)

- Faux négatif : terme candidat valide et non extrait par l'ETC.
- Précision (sensibilité ou exactitude) :  $VP / (VP + FP)$ . Elle permet d'évaluer le taux de termes candidats valides et extraits par l'ETC, en comparaison au nombre total de termes candidats extraits (valides et non valides) par l'ETC. La précision donne une information sur le bruit produit par l'ETC. Plus la précision est basse, plus le bruit (TC extraits incorrectes) est important.
- Rappel (sélectivité) :  $VP / (VP + FN)$ . Il permet d'évaluer le nombre de termes candidats valides et extraits par l'ETC, en comparaison au nombre de termes candidats valides et non extraits. Il donne une marge d'erreur. Plus le rappel est élevé, plus le silence est bas, ce qui revient à dire que l'ETC a identifié les termes candidats valides et les a extraits.
- F-mesure :  $2 * P * R / (P + R)$ . Donne une note globale, calculée par la combinaison pondérée de la précision et du rappel.

TABEAU 3.2 – Résultats de l'analyse du corpus test par les ETC.

Nombre de terme candidat ...	YATEA	BIOTEX
...extraits	751	1275
...vrais positifs (VP)	311	472
VP adéquates à la validation manuelle (528)	269	323
VP obtenus par dérivation automatique	13	71
Autres VP valides	<b>29</b>	<b>78</b>
...faux positifs	440	803
...faux négatifs	217	56
Précision	<b>41</b>	37
Rappel	59	<b>89</b>
F-mesure	48	<b>52</b>

Les résultats indiquent que ces deux ETC proposent beaucoup de bruit (réponses fausses) dans les résultats. Cela explique le travail manuel très important qui est généré par les extracteurs. En effet, dans le cas de YATEA, 59% des réponses proposées sont fausses et pour BIOTEX c'est pire avec 63% de réponses fausses. Cependant, le taux de rappel de BIOTEX est bon, seul 11% de termes candidats validés manuellement n'ont pas été identifiés. Dans le cas de YATEA c'est 41% de silence, de termes candidats valides que le logiciel a laissé de côté. La F-mesure confirme ces résultats et juge légèrement plus performant BIOTEX en raison de son très bon taux de rappel. De plus, BIOTEX a permis d'identifier 78 termes qui n'avaient pas été identifiés lors de la validation manuelle, alors que YATEA n'en propose que 29 de plus. En outre, nous constatons que les résultats obtenus par YATEA correspondent à l'évaluation qu'avait réalisé [Alecu et al. \[2012\]](#).

### 3.3 Développement du modèle par approche descendante (top-down)

Les méthodes descendantes se concentrent sur la réutilisation de modèles génériques [[Aussenac-Gilles et Charlet, 2010](#)]. C'est donc la gamme de modèles déjà existants qui va guider la modélisation, pour répondre aux besoins applicatifs. En outre, les modèles

génériques sont développés avec un haut niveau d'abstraction, afin de leur permettre de s'imbriquer dans toute application [Charlet et Bachimont, 1998]. La figure 3.5 illustre l'approche qui va d'une bibliothèque de modèles à un modèle adapté, via l'intervention au minimum d'un ontologue et d'un acteur du domaine.

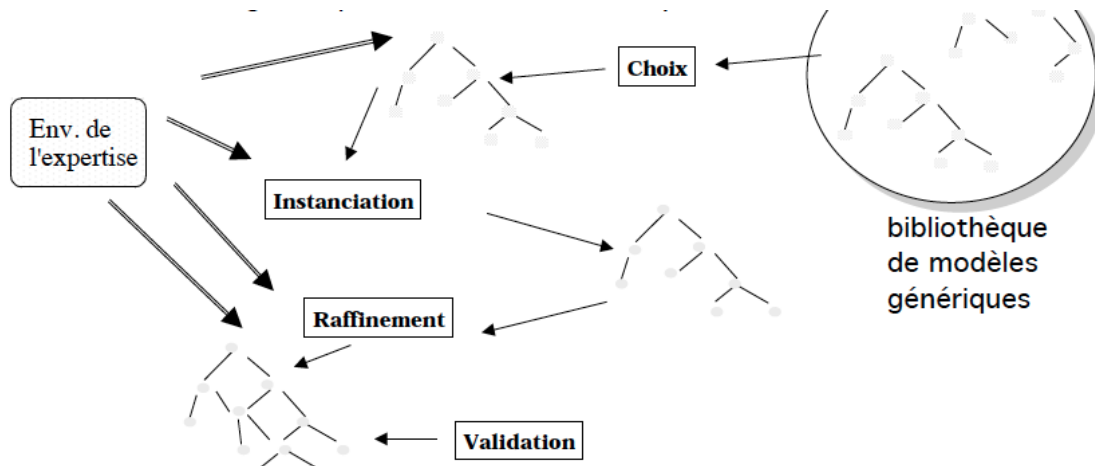


FIGURE 3.5 – De modèles génériques vers un modèle adapté [Aussenac-Gilles et Charlet, 2010].

### 3.3.1 Le projet Common Kads

Le projet européen Common Kads, initialement Kads, a démarré en 1983 et vise à l'élaboration d'une approche structurée, pour le développement de systèmes à base de connaissances (SBC) [Schreiber *et al.*, 1994]. La méthodologie a été affinée au fil des années et validée par des partenaires universitaires (dont le principal est l'Université d'Amsterdam) et commerciaux dans le cadre du Programme Européen ESPRIT. Le projet propose de modéliser l'ensemble des connaissances à l'aide d'un ensemble de modèles génériques. L'utilisateur sélectionne son modèle via sa définition au sein d'une bibliothèque de modèles. Ce modèle est ensuite instancié avec les connaissances du domaine à modéliser. Cette méthodologie repose en partie sur l'idée que la connaissance humaine a une structure stable et que cette structure peut être rendue générique, par la construction de modèles représentant ses différents aspects [Schreiber, 2000].

- Le modèle organisationnel :
  - permet de définir les besoins, le cadre, et la faisabilité du SBC.
  - but : identifier les problèmes, les opportunités et les impacts potentiels du développement d'un SBC.
- Le modèle de tâche : permet de décrire les tâches et sous tâches qui seront réalisées par l'agent.
- Le modèle agent : permet de décrire les capacités, normes, préférences et permissions accordées à l'agent en tant qu'exécuteur de tâche (un humain ou une machine).
- Le modèle de connaissances : « fournit la description conceptuelle des méthodes de résolution de problèmes et des données qui doivent être traitées et livrées par le système. » [Schvartz *et al.*, 2007]

### 3.3 Développement du modèle par approche descendante (top-down)

- Le modèle de communication : permet de décrire les interactions entre les agents (humains ou machines) sous forme d'acte du langage.
- Le modèle de conception : permet de décrire la structure du système qui doit être développé.

Le projet propose également une typologie des modèles de connaissances : domaine, inférence et tâche répertorié dans un modèle à trois couches illustré figure 3.6

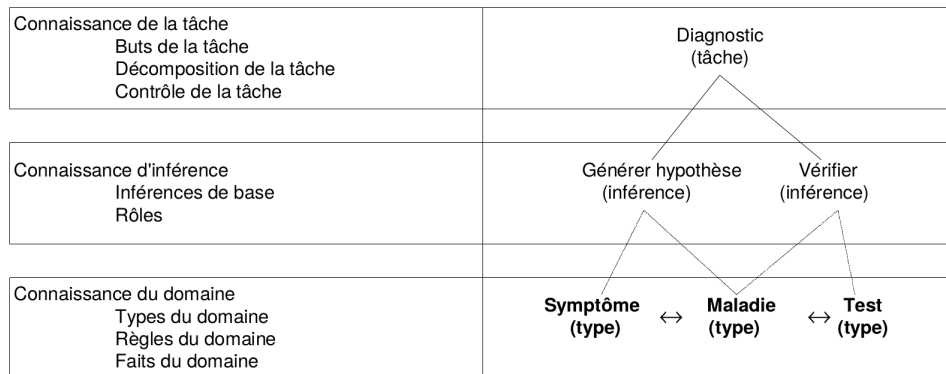


FIGURE 3.6 – Le modèle de connaissances de COMMONKADS d'après Schreiber [2000].

- La connaissance du domaine : « permet de décrire les types de connaissances propres à l'application tels que les données, les concepts, les hiérarchies de concepts, les relations, et les règles de résolution de problèmes. » [Schvartz *et al.*, 2007]
- La connaissance d'inférence : « permet de produire de nouvelles connaissances à partir de connaissances existantes. Une inférence représente une étape élémentaire du processus de raisonnement. » [Schvartz *et al.*, 2007]. Les entrées et sorties des inférences sont des connaissances du domaine appelées des rôles. Par exemple, dans une tâche de diagnostic, l'inférence "générer une hypothèse" associe des symptômes à une maladie et l'inférence "vérifier" permet d'identifier les tests qui infirment ou confirment que les symptômes ont été causés par la maladie.
- La connaissance de la tâche : « décrit les buts poursuivis dans une application donnée et les moyens d'atteindre ces buts en définissant la façon dont les inférences sont utilisées. Une tâche est décomposée en sous-tâches et ultimement en inférences » [Schvartz *et al.*, 2007]. Une tâche peut être ainsi décrite pour plusieurs domaines. Le diagnostic médical ou le diagnostic d'une panne sur une machine quelconque font appel au même raisonnement : "générer une hypothèse" et "vérifier", ce sont les rôles propres à chaque domaine qui change.

#### 3.3.2 Le projet Sensus

Le projet SENSUS est mené par The Natural Language Group (NLG) au sein de l'Information Science Institute (ISI) de l'Université de Californie du Sud. Ce projet propose une large ontologie regroupant actuellement près de 90 000 concepts, pour développer des ontologies de domaine de manière semi-automatique [Swartout *et al.*, 1996]. L'ontologie a été développée en suivant la méthodologie décrite dans Knight et Luk [1994]. Les

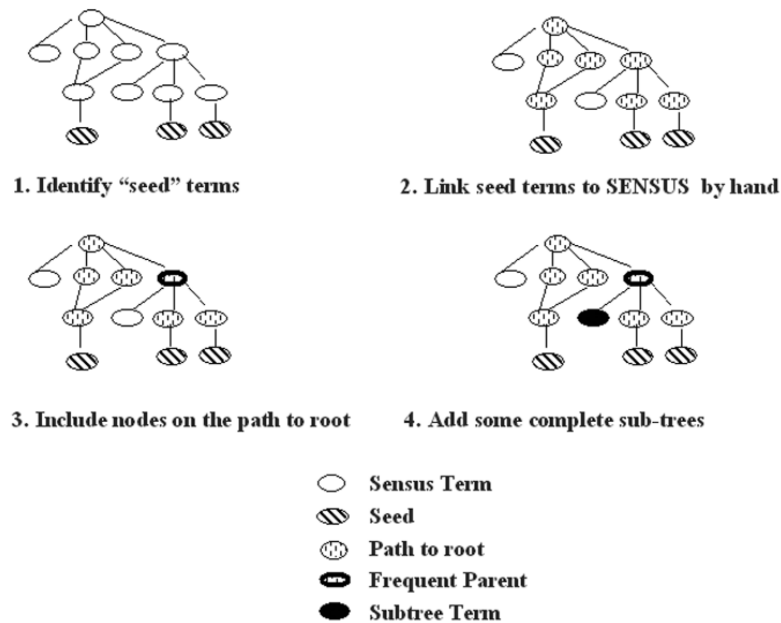


FIGURE 3.7 – La méthodologie de SENSUS illustrée dans Swartout *et al.* [1996].

auteurs ont fusionné différentes ressources existantes pour créer une ontologie de base, telles que la top ontologie PENMAN Upper Model<sup>13</sup> ou la base lexicale WordNet<sup>14</sup> et des dictionnaires pour assurer les tâches de traduction. Les concepts sont organisés selon leur niveau d'abstraction (voir en section 2.1.3). Un exemple d'utilisation de SENSUS appliqué au développement d'une ontologie de domaine de la planification de la campagne aérienne militaire est proposé dans Swartout *et al.* [1996] et illustré figure 3.7 :

1. Sélection par un expert des concepts de domaine de SENSUS réutilisables dans l'ontologie (1. de 3.7).
2. Inclusion automatique de tous les concepts dans le chemin allant du terme sélectionné à la racine de SENSUS (2. de 3.7).
3. Ajout manuel de termes adéquates au domaine tels que les termes sous les chemins précédemment inclus (3. de 3.7).
4. Inclusion automatique de branches entières de l'arbre de l'ontologie à partir des termes ajoutés manuellement (4. de 3.7).

Cette méthode met donc l'accent sur la réutilisation de modèles, étant donné qu'une seule et même ontologie permet de construire un ensemble d'ontologies de domaine. Nous notons également que cette méthodologie est implémentée sous le web serveur Ontosaurus également développé par ISIS.

13. Cette ontologie a été développée dans le but de servir les applications basées sur le traitement du langage : <http://www.fb10.uni-bremen.de/anglistik/langpro/kpml/um89/um89-root.htm>

14. <http://wordnetweb.princeton.edu/perl/webwn>

## 3.4 Développement du modèle par approche hybride (bottom-up et top-down)

Tel qu'il est dit par [Charlet et Bachimont \[1998\]](#), les méthodes de constructions tendent naturellement vers une des deux approches présentées : soit vers les approches descendantes, quand la réutilisation est fortement recherchée ; soit vers les approches ascendantes, quand c'est le développement d'un nouveau modèle qui est souhaité. La majorité des méthodes sont donc hybrides. En effet, un modèle construit par méthode ascendante peut très bien être enrichi avec un modèle existant. Et vice versa, un modèle existant peut permettre d'engager une modélisation. Nous pouvons rappeler en outre que les chercheurs en IC s'accordent à dire que la méthodologie choisie dépend principalement du but visé par la modélisation et des données disponibles pour la construction.

### 3.4.1 Macao

MACAO est un outil d'acquisition de connaissances développé dans le cadre de la thèse de Aussenac-Gilles [[Aussenac-Gilles, 1989, 2005](#)]. L'outil est indépendant de la tâche et du domaine. La méthode est partisane d'une plus grande collaboration entre les ontologues et les experts du domaine, qui devraient prendre part activement au processus de développement des ontologies de domaine.

Le but de cette méthode est d'extraire les connaissances d'un domaine sans préjugé de l'utilisation que l'ontologue pourra en faire. L'étape 1 permet de survoler l'ensemble du domaine à modéliser, pour en comprendre les enjeux et problématiques. Les étapes 2 et 3 visent à l'obtention de connaissances du domaine, en démarrant la modélisation conceptuelle. L'étape 2 suit une approche bottom-up, avec le développement d'un premier cadre conceptuel. L'étape 3 suit une approche top-down, dans le but d'affiner le modèle et de le compléter, par le biais de modèles déjà existants. Enfin, la dernière étape vise à l'opérationnalisation de l'ontologie dans un SBC afin de la valider.

### 3.4.2 ToReuse2Onto

Cette méthode a été développée dans le cadre de la thèse de [Drame \[2014\]](#). Elle reprend les bases de la méthodologie TERMINAE, avec une approche multilingue réalisée par un module d'alignement de termes, et la réutilisation de RTO effectuée dès la phase de conceptualisation. TOREUSE2ONTO se décompose en cinq étapes [[Drame, 2014](#)] :

1. **La constitution du corpus** : qui doit répondre aux exigences de couverture du domaine et d'adéquation aux tâches d'extraction automatique de termes et relations.
2. **L'extraction des candidats termes** : est réalisée à l'aide de méthodes et outils du TAL développés pour l'acquisition de termes à partir de textes (voir en section [3.2.3](#)).
3. **La construction du noyau de l'ontologie** : est l'étape de construction de l'arborescence conceptuelle, à partir des termes extraits à l'étape précédente. L'auteur conseille de réaliser cette étape à l'aide de ressources sémantiques existantes. Ces dernières permettent de faire un alignement avec les termes extraits pour cibler les concepts.

4. **L'enrichissement de l'ontologie** : a pour but l'intégration de nouveaux artefacts, qui peuvent être des concepts ou des relations et les termes associés. Pour l'intégration de nouveaux concepts se sont les dépendances syntaxiques qui sont analysées. Ces dernières permettent d'inférer une relation taxonomique entre deux termes. Par exemple, les termes « schizophrénie », « schizophrénie affective » et « schizophrénie paranoïde » de part leur tête syntagmatique commune ont de forte chance d'être des concepts reliés par une relation taxonomique. Pour trouver de nouveaux termes, il est possible de réaliser un alignement de termes de langues différentes, basé sur des corpus parallèles.
5. **La validation et la formalisation de l'ontologie** : permet de vérifier et valider l'ontologie. Cette méthode se fonde sur le savoir des acteurs du domaine modélisé.

Cette méthodologie a été mise en œuvre dans le cadre du développement d'une ontologie sur la maladie d'Alzheimer [Dramé *et al.*, 2014].

### 3.5 Discussions sur les méthodes de construction d'ontologies

Les méthodologies s'articulent autour de deux axes : ascendant et descendant. Le premier offre un cadre méthodologique pour le développement d'ontologies à partir de données brutes. Que les méthodes ascendantes soient manuelles ou semi-automatisées, elles sont souvent coûteuses et fastidieuses à mettre en œuvre. Elles demandent une intervention humaine importante. Les outils qui permettent d'extraire les connaissances automatiquement sont venus apporter un soutien pour l'acquisition des connaissances. Toutefois, ces outils nécessitent des ressources linguistiques importantes, dans une langue adaptée à l'outil, ainsi qu'une expertise manuelle complexe des connaissances extraites. On observe également que ces méthodes de construction d'ontologies ont engendré un effacement progressif des acteurs du domaine et placé l'ontologue au centre du processus. Cependant, ces acteurs du domaine demeurent les détenteurs de la connaissance encyclopédique et pratique qui peut faire défaut à l'ontologue. On observe également que ces méthodes automatiques ont permis de développer des ontologies de taille plus importante, entraînant du même coup une plus grande difficulté à assurer une modélisation adéquate et correcte. La validation d'ontologies est par conséquent devenue une problématique à part entière de l'ingénierie des connaissances (IC).

En parallèle, les approches descendantes offrent des solutions méthodologiques, pour la réutilisation de modèles déjà existants. Ces méthodes présentent de nombreux freins. La réutilisation d'un modèle impose d'en adopter son formalisme et l'incomplétude de ces modèles de base peut, in fine, demander autant d'effort de recueil de connaissances ou d'expertise manuelle, que le développement d'un modèle à partir de données brutes. En outre, ces méthodes ont participé à la naissance d'un intérêt particulier pour le développement modulaire, qui permet de s'approprier plus facilement des ontologies déjà existantes.

Enfin, des méthodes plus récentes ont vu le jour, elles proposent l'utilisation de ces deux approches. Ces méthodes hybrides tiennent compte des modèles existants et misent sur leur appropriation et leur enrichissement, par l'ajout de nouvelles connaissances. Selon les domaines d'application, ces méthodes peuvent s'avérer très efficaces. Par exemple, en biomédical, les classifications et thésaurus sont très présents et il sont des systèmes



d'organisation des connaissances déjà validés par la communauté scientifique. De plus, le partage des ontologies biomédicales est facilité par les plateformes dédiées telles que BioPortal <sup>15</sup>.

Toutefois, des difficultés persistent, notamment concernant les outils d'extraction automatique de connaissances. Ces outils basés sur des techniques linguistiques ou/et de statistiques demandent encore une expertise manuelle très importante. Nous avons expérimenté cette contrainte dans nos travaux présentés au chapitre 5. De plus, leur utilisation est contrainte par la langue du corpus. Ensuite, la réutilisation de modèles existants, bien que plus rapide à mettre en œuvre et permettant d'obtenir des résultats plus prévisibles présente aussi des limites. En effet, le choix du modèle de départ contraint fortement l'ontologie résultante. Nous avons également expérimenté cette contrainte dans nos travaux présentés au chapitre 6.

## 3.6 La modularité ontologique : l'ergonomie au service du développement du modèle.

Pour conclure ce chapitre consacré à la construction d'ontologies, nous présentons la modularité ontologique. Elle est une des solutions permettant de gérer efficacement les ontologies : leurs réutilisations, leurs maintenances, leurs personnalisations entre autres. Nous avons nous-mêmes, dès le début de notre projet, choisi une approche modulaire pour modéliser différents sous-domaines propres à la psychiatrie.

### 3.6.1 Définition de la modularité ontologique

La modularité ontologique se définit sous différents aspects. Selon les définitions suivantes, un module est une ontologie indépendante qui peut-être intégrée à une ontologie d'ensemble, dans laquelle il conserve ses propriétés. Ainsi, l'ontologie d'ensemble sert de lien entre les modules. Par exemple, dans notre projet, ONTOPSYCHIA se définit par l'assemblage de trois modules ontologiques qui conservent leurs propriétés et leur indépendance.

**Définition de Rector [2002] :** « Modularité : les différentes parties d'une ontologie qui peuvent être construites séparément puis ensuite combinées sans causer une explosion combinatoire. »

**Grau et al. [2006] et Doran [2006]** ajoutent qu'un module ontologique est caractérisé par la possibilité d'être lui-même modularisé ; augmentant d'autant plus les possibilités de réutilisation des modules. Un module ontologique est donc une unité, qui est à la fois indépendante et intégrable à une ou plusieurs ontologies. Plusieurs modules d'ontologies peuvent être ainsi combinés pour coexister au sein de larges ontologies.

---

15. <http://biportal.bioontology.org/>



**Définition de Cuenca Grau et Kutz [2007] :** « Dans une approche formelle, un module est représenté comme un composant d'un tout (avec son propre langage, et sa propre sémantique) ». Ils précisent en outre, que le modèle engendré par l'interaction de modules produit une nouvelle syntaxe, avec sa propre sémantique.

**Définition Pathak *et al.* [2009] :** « Un module est un sous-ensemble d'un tout, qui est porteur de sens (i.e., il n'est pas un sous-ensemble arbitraire construit aléatoirement) et il peut exister indépendamment du tout, bien qu'il ne supporte pas obligatoirement les mêmes fonctionnalités que le tout. »

« Un module ontologique est un composant réutilisable d'une plus large et plus complexe ontologie, qui est autonome, mais porte une association définie à d'autres modules d'ontologies, incluant l'ontologie originelle. »

Ces définitions introduisent la notion de dimension d'un module ontologique, par rapport à une ontologie modulaire. En effet, si un module est une partie d'ontologie modulaire, alors une ontologie composée de différents modules est plus large que les modules en eux-mêmes.

### 3.6.2 Objectifs de la modularité ontologique

La réutilisation, ainsi que la maintenance et la gestion des ontologies étaient des points cruciaux abordés dès le développement de l'ingénierie des connaissances. Cependant, victime de sa popularité, l'IC a dû adapter rapidement les techniques de la modularité [Del Vescovo *et al.*, 2013], afin de satisfaire ces engagements applicatifs. La modularité des ontologies est vue depuis comme « l'exigence clé » pour les grandes ontologies, afin de parvenir à « la réutilisation, la maintenance, et l'évolution » [Rector, 2003]. Elle permet également une meilleure compréhension, facilite le raisonnement ou la création d'extensions [Cuenca Grau et Kutz, 2007].

Depuis 2006, se tient chaque année un Workshops<sup>16</sup> sur les ontologies modulaires. Il est l'occasion de discuter des défis liés à la modularité, tels que celui de l'interrelation entre des ontologies couvrant le même domaine de connaissances. En effet, une ontologie couvrant un certain domaine va seule être consistante, mais la jointure de plusieurs ontologies peut donner lieu à des inconsistances [Del Vescovo *et al.*, 2013]. Ce Workshops est aussi l'occasion d'aborder les objectifs majeurs de la modularité : permettre la réutilisation des ontologies, ainsi que leur maintenance et gestion, entre autres [Rector, 2003, Cuenca Grau et Kutz, 2007, Pathak *et al.*, 2009, Del Vescovo *et al.*, 2013].

En IC, la notion de modularité a pu être objectivée en mettant en avant les différents points qu'elle vise à améliorer. [Spaccapietra *et al.*, 2005] :

- **L'interrogation des données et le raisonnement :** les modèles ontologiques peuvent posséder un nombre illimité de classes, de relations, d'objets, de règles et d'axiomes. Cet objectif introduit donc le problème de la complexité des données qu'un système de requête doit parcourir afin de fournir une réponse. La modularité permet de gérer des ensembles de connaissances plus petits et plus rapides à parcourir, si l'organisation sémantique est bien gérée au sein des modules.

16. <http://iaoa.org/womo/>

- **La réutilisation** : certainement l'une des ambitions les plus prisées dans le développement d'ontologies. Plus un module sera compréhensible et facile à gérer, plus il aura de chance d'être réutilisé et intégré à de nouvelles applications.
- **L'appropriation et la personnalisation** : sont en lien avec la réutilisation. La modularité permet de s'approprier un module ontologique déjà existant, pour alimenter une ontologie en développement. La personnalisation offre la possibilité de répondre à des besoins de modélisations spécifiques.
- **La gestion de la complexité** : plus les modèles sont grands, plus il est difficile de garantir la qualité et l'exactitude des connaissances qu'ils modélisent. Pour comprendre cet objectif, le plus simple est de se représenter qu'un travail sur des modules à « taille humaine » permet d'en faciliter la gestion par l'humain.
- **La compréhension** : est proche de la gestion de la complexité. L'objectif étant de se donner les moyens de comprendre facilement et rapidement l'ontologie développée. Si le contenu visualisé est de taille réduite, la compréhension en sera d'autant plus accessible.
- **L'évolution et la maintenance des modèles ontologiques** : les connaissances évoluent, donc les modèles qui les représentent aussi. Les ontologies doivent être mises à jour avec l'évolution des connaissances des domaines qu'elles modélisent. La mise à jour d'un type de données doit pouvoir se répercuter aisément à l'ensemble de l'ontologie. La gestion en modules permet de réduire les coûts de maintenances, en ne modifiant qu'un module dont les changements vont se répercuter sur l'ensemble du modèle.

Nous observons également dans la littérature, que la modularité en IC ne s'articule pas seulement autour de nouvelles méthodes de construction d'ontologies. La question de l'utilisation des grandes ontologies déjà développées se pose aussi au travers de la modularité. Ces ontologies sont parfois difficilement exploitables en leur totalité. Leur réutilisation peut donc passer par leur fractionnement en modules. Un exemple récent est celui engendré par le programme GALEN [Seidenberg et Rector, 2006].

#### 3.6.3 La composition modulaire

Cette méthode se calque sur les principes de l'intégration d'ontologies. Les modules, construits indépendamment les uns des autres selon des thèmes qui leur sont propres, sont ensuite assemblés pour former une grande ontologie. Pinto et Martins [2000] : « la construction peut se faire par assemblage, extension, spécialisation ou adaptation, d'autres ontologies, qui s'intègrent à l'ontologie résultante/finale. Ou par la fusion de différentes ontologies abordant le même thème et qui en résulte une ontologie unifiée. »

**Plugin Protégé 4.0**<sup>17</sup> pour l'importation et l'intégration d'ontologies. La dernière version de Protégé offre une facilité de gestion de la modularité grâce à un import intelligent des modules ontologiques. En effet, à chaque chargement de module dans un même espace

---

17. <http://protegewiki.stanford.edu2009>

de travail, Protégé va fusionner les différents concepts pour ne former qu'une grande ontologie. Les modules sont unifiés au sein d'une nouvelle ontologie. Ils perdent donc leur fonction de module au sein de cette nouvelle ontologie.

**NeOnMetamodel** a lui aussi développé une suite de plugin pour la gestion d'ontologies modulaires. [D'Aquin et al. \[2008\]](#) font un état des lieux des développements de l'éditeur d'ontologies en matière de modularité. À noter qu'il gère à la fois, la composition et la décomposition.

**ANEMONE Ontology Development Methodology** tend à définir une méthodologie de développement d'ontologies modulaires, en plusieurs étapes. Cette méthodologie est explicitée dans un article servant de guide de développement [\[Özacar et al., 2011\]](#). Avant tout développement, la méthodologie pose une typologie des modules, correspondant à la spécificité des concepts qui seront modélisés. Du plus large au plus spécialisé nous avons : les modules ontologiques de base, les modules ontologiques de domaine d'ordre supérieur, les modules ontologiques de domaine, et les modules ontologiques locaux. Cette hiérarchie n'est pas sans rappeler la typologie des ontologies divisées respectivement en : Top-ontologie, core-ontologie, et ontologie de domaine (présentée en section [2.1.3](#)). Les modules ontologiques locaux permettent deux choses : (1) redéfinir ou renommer un concept de domaine sans le modifier à la source, (2) conserver un contenu assertif sur des concepts de modules ontologiques de domaine.

**The Distributed ontology, Modelling and Specification Language (DOL)** [\[Lange et al., 2012\]](#) propose un métalangage pour unifier (1) des ontologies formalisées dans des logiques hétérogènes, (2) des ontologies modulaires, (3) des liens entre les ontologies et (4) l'annotation des ontologies. Ce niveau de métalangage permet d'établir une interopérabilité conceptuelle entre différentes ontologies.

### 3.6.4 La décomposition modulaire

La décomposition va créer des modules à partir d'une grande ontologie [\[Pathak et al., 2009\]](#). L'approche se veut semi-automatique. Les modules obtenus sont ensuite validés par un expert. Cette approche pose généralement le problème de la complétude. En effet, il ne faut pas perdre d'informations après un découpage en modules. Une requête doit donc fournir le même résultat avant et après la décomposition.

**Le projet GALEN** a débuté en 1999, suite au programme GALEN. Ce dernier a été impulsé par Alan Rector afin de développer une terminologie médicale. Le programme GALEN a permis la création d'une large ontologie du domaine médical [\[Rector et al., 2003\]](#). Quelques années après le démarrage du programme, l'ontologie contenait déjà quelques milliers de concepts, qui rendaient son utilisation et sa maintenance difficile [\[Seidenberg et Rector, 2006\]](#). Un outil de segmentation de l'ontologie en module a ainsi été développé. Il s'appuie en entrée sur un ou plusieurs noms de concepts donnés par l'utilisateur. Il va ensuite effectuer un balayage vertical pour récupérer l'arborescence des concepts l'entourant, de la racine à la dernière feuille. Il va également suivre tous les chemins transversaux qui partiront de ces concepts. L'algorithme se termine quand il ne trouve plus de

chemins à suivre, ce qui signifie que tous les concepts dans le voisinage du concept en entrée sont extraits.

## 3.7 Synthèse

Les ontologies sont des objets complexes qui demandent la maîtrise de technologies spécifiques à leur développement. La construction d'une ontologie oblige à faire preuve de méthode, car elle est un projet en tant que tel, soumise à un cahier des charges et un temps de développement.

Dans ce chapitre, nous avons exploré les propositions méthodologiques importantes qui ont vu le jour pour l'aide à la construction d'ontologies. Nous avons vu qu'il est nécessaire de procéder par étape : évaluation des besoins, recueil des données, développement du modèle et validation de l'ontologie. Les méthodes s'articulent autour de deux axes et répondent ainsi à deux problématiques. Les approches ascendantes (manuelles ou semi-automatiques) offrent des cadres méthodologiques pour le développement d'ontologies à partir de données brutes. Les méthodes entièrement manuelles ont l'avantage de ne pas être dépendantes de la langue ; alors que les méthodes par extraction de termes sont contraintes par la langue du corpus. Les approches descendantes offrent des solutions méthodologiques, pour la réutilisation de modèles déjà existants. Elles ont l'avantage de ne pas partir de zéro, mais elles nécessitent d'adopter le formalisme utilisé dans le modèle ou dans le cas de COMMON KADS de se former à l'instanciation des modèles de base. Enfin, les méthodes hybrides, qui mixent l'utilisation de ces deux approches, tiennent compte des modèles existants et misent sur l'appropriation et l'enrichissement de ces modèles. Ce type d'approche est particulièrement adapté à des domaines comme le biomédical, qui contiennent déjà un grand nombre de classifications et thésaurus validés par la communauté scientifique. Enfin, La modularité est également un des outils majeurs pour faciliter le développement, la réutilisation et la maintenance des ontologies.

La dernière étape au développement d'une ontologie est sa validation. Nous abordons cette problématique dans le chapitre suivant.



# Chapitre 4

## L'art de valider une ontologie

### Sommaire

<b>4.1 La définition des critères de validation</b>	<b>86</b>
4.1.1 Les critères de Gruber [1995]	86
4.1.2 Les critères de Gómez-Pérez <i>et al.</i> [1995]	87
4.1.3 Les critères de Guarino et Welty [2000]	88
4.1.4 Les critères de Poveda-Villalón <i>et al.</i> [2012]	88
4.1.5 Des définitions standardisées aux problèmes de qualité des ontologies, selon les travaux de Gherasim <i>et al.</i> [2012]	89
4.1.6 Synthèse	90
<b>4.2 Validation de la structure</b>	<b>90</b>
4.2.1 Les moteurs d'inférence	90
4.2.2 Les méthodes et outils	91
<b>4.3 Validation de la sémantique</b>	<b>93</b>
4.3.1 Un développement collaboratif couplé à des approches ergonomes en réponse aux difficultés liées à la validation sémantique	93
4.3.2 Les outils collaboratifs pour la validation de la sémantique	94
4.3.3 Les limites du « tout collaboratif » pour la validation de la sémantique	94
<b>4.4 Synthèse</b>	<b>94</b>

Nous venons d'approfondir, au travers des chapitres précédents, la notion de modélisation des connaissances à l'aide des ontologies informatiques. Ces dernières années, nous avons observé un développement croissant des méthodes de construction d'ontologies fondées sur l'extraction de termes spécialisés au sein de corpus. Nous observons également que ces méthodes automatiques ont permis de développer des ontologies plus volumineuses, entraînant du même coup une plus grande difficulté à assurer une modélisation adéquate et correcte. Une problématique majeure reste donc en suspens : la validation des ontologies. En effet, qu'est-ce qu'une ontologie valide ? Quels sont les critères qui garantissent la validité d'une ontologie ? Existence-ils des méthodes pour valider les ontologies ? Dans un ouvrage dédié à la validation d'ontologies, [Vrandečić \[2009\]](#) a relevé trois scénarios qui justifient la validation, nous les résumons comme tels : (1) une ontologie validée permet une meilleure réutilisation des données modélisées ; (2) les ontologues qui disposent de méthodes pour évaluer et valider leurs modèles peuvent partager leurs résultats avec la communauté, et également réutiliser avec confiance le travail des autres à leurs propres fins ; (3) les coûts d'entretien des ontologies sont moindres, quand les méthodes d'évaluation d'ontologies permettent de vérifier automatiquement si les contraintes et les exigences sont remplies. Nous notons également dans la littérature que la validation se définit sous deux aspects complémentaires : (1) la validation structurelle et logique, qui peut être réalisée automatiquement, grâce au développement d'outils dédiés et (2) la validation sémantique et sociale, qui peine encore à trouver des méthodes consensuelles. Dans ce chapitre, une première section vise à définir ce qu'est une ontologie validée, à l'aide de listes de critères de validation qui objectivent la qualité de l'ontologie. Une deuxième section fait un point sur les méthodes de validation de la structure. Et une troisième section explore les approches pour la validation de la sémantique.

## 4.1 La définition des critères de validation

Avant même que les méthodes de validation d'ontologies ne commencent à se développer, de nombreux chercheurs se sont intéressés aux critères de validation qui permettent d'objectiver la qualité attendue d'une ontologie. Nous présentons dans cette section, quatre grandes listes de critères. [Sabou et Fernandez \[2012\]](#) reprennent dans un ouvrage dédié à la validation d'ontologies, les grandes méthodes de validation qui se sont développées et les critères pris en compte. Ils notent également que les critères qui permettent de valider une ontologie dépendent bien souvent de son utilisation finale.

### 4.1.1 Les critères de [Gruber \[1995\]](#)

[Gruber \[1995\]](#) indique que lorsque nous choisissons de représenter quelque chose dans une ontologie, nous faisons par là même un choix de représentation, de modélisation. Et ce choix doit être guidé par des critères objectifs, en accord avec le but visé par l'artefact. L'auteur propose alors une liste de critères pour aider au développement d'une ontologie valide. Et, en particulier pour le développement des ontologies, dont le but est le partage des connaissances et l'interopérabilité entre programmes informatiques basés sur une conceptualisation commune.

1. **La clarté** : les définitions des concepts doivent être objectives. Une définition complète (nécessaire et suffisante) est à préférer à une définition partielle. Et chaque

#### 4.1 La définition des critères de validation

---

définition doit être documentée en langage naturel.

2. **La cohérence** : une ontologie ne doit autoriser que les inférences en adéquation avec les définitions.
3. **L'extension** : l'ontologie doit être pensée en anticipant les utilisations du vocabulaire partagé. Il doit être possible de définir de nouveaux termes, pour une utilisation nouvelle, en se basant sur le vocabulaire et les définitions déjà intégrés à l'ontologie.
4. **Un biais d'encodage minimal** : la conceptualisation doit être spécifiée au niveau des connaissances, sans dépendre d'un encodage particulier. C'est-à-dire que les choix de modélisation ne doivent pas dépendre d'un système d'implémentation spécifique ou de notations particulières.
5. **Un engagement ontologique minimal** : l'ontologie doit contenir aussi peu que possible d'affirmation sur le monde modélisé. Le but est de permettre à l'ontologie d'être instanciée et spécialisée à souhait.

L'engagement ontologique est discutable selon le but visé par l'ontologie. En effet, la construction d'une ontologie de domaine nécessite de faire des choix. Par exemple, dans une ontologie de la psychiatrie nous faisons des affirmations sur le classement des troubles.

##### 4.1.2 Les critères de **Gómez-Pérez et al.** [1995]

**Gómez-Pérez et al.** [1995] discutent les ressemblances et les différences entre les bases de connaissances et les ontologies. Le but de cette comparaison est d'établir une corrélation pour la validation d'ontologies, à partir des méthodes de validation déjà existantes pour les bases de connaissances. Leurs travaux ont permis d'établir une liste de critères pour la validation d'ontologies, qui a pu être expérimentée dans la méthode de construction d'ontologies METHONTOLOGY (voir en section 3.2.1 [**Gómez-Pérez**, 2001]) :

1. **La consistance** : fait référence au fait d'obtenir des conclusions contradictoires à partir de définitions valides. Une définition est ainsi consistante si et seulement si aucune phrase contradictoire ou aucun sens contradictoire ne peuvent en être inférés, directement ou indirectement (notamment par la biais d'autres définitions et axiomes qui la définissent). Par exemple, si une ontologie modélise les humains par l'union disjointe des hommes et des femmes, mais que la même ontologie modélise les hermaphrodites comme des humains à l'intersection des hommes et des femmes, il y a une contradiction, donc le modèle est inconsistant.
2. **La complétude** : elle ne peut pas être démontrée. Mais il est possible de prouver l'incomplétude (a) d'une définition individuelle et d'en déduire alors l'incomplétude de l'ontologie ou (b) si une définition est manquante dans le cadre de référence établi, nous prouvons l'incomplétude de l'ontologie. En d'autres termes, une ontologie est complète si et seulement si : (1) tout ce qui est supposé être dans l'ontologie est explicitement exposé dedans, ou peut être inféré ; (2) toutes les définitions sont complètes.
3. **La concision** : une ontologie est concise si (a) elle ne contient aucune définition non-nécessaire ou inutile, (b) il n'existe pas de redondance explicite entre les définitions de termes, (c) des redondances ne peuvent pas être inférées à partir d'autres définitions et axiomes.



4. **L'évolutivité** : elle est garantie s'il est possible d'ajouter de nouvelles définitions et plus de connaissances, sans altérer l'ensemble des propriétés déjà correctement définies.
5. **La sensibilité** : elle se réfère à la manière dont de petits changements dans une définition altèrent l'ensemble des propriétés déjà formellement définies.

#### 4.1.3 Les critères de **Guarino et Welty [2000]**

**Guarino et Welty [2000]** énoncent également des critères, basés sur des principes philosophiques déjà existants, qui visent à définir un cadre permettant d'affirmer que la taxonomie d'une ontologie est valide. Toutefois, cette méthode ne permet pas de définir que la modélisation mise en œuvre dans l'ontologie est adéquate avec le domaine. Ces critères ont été implémentés dans la méthode ONTOCLEAN. Nous définissons brièvement ces critères, des définitions plus exhaustives ainsi que des exemples concrets sont développés dans **Guarino et Welty [2009]**.

1. **L'essence et la rigidité** : « La propriété d'une entité est essentielle à cette entité si elle doit être vraie dans tous les mondes possibles. [...] Une propriété est rigide si elle est essentielle à toutes les instances possibles. [...] La rigidité est une méta-propriété qui permet de décider si les propriétés d'une ontologie sont pertinentes [...] Il existe également des propriétés non-rigides, qui peuvent acquérir et perdre leurs instances selon un état donné. Nous distinguons alors les propriétés non essentielles à certaines entités (*semi-rigide*) et les propriétés qui ne sont essentielles à aucune de leur instance (*anti-rigide*). ». Les auteurs insistent sur l'importance de la rigidité et appuient que toutes les propriétés d'une ontologie devraient être annotées en « rigide », « non-rigide » ou « anti-rigide ». Car, « en plus de fournir une information supplémentaire sur la signification attendue d'une entité, ces méta-propriétés imposent des contraintes sur la relation de subsomption. Contraintes qui permettent par la suite de vérifier la consistance logique des liens d'une taxonomie ». Par exemple, l'anti-rigidité ne peut subsumer la rigidité. Ainsi, le fait d'être étudiant (propriété anti-rigide) ne peut subsumer le fait d'être humain (propriété rigide).
2. **L'identité et l'unité** : sont pour les auteurs les notions les plus importantes de leur méthodologie. L'identité renvoie à la problématique de reconnaître des entités individuelles du monde comme étant les mêmes. Et l'unité renvoie à la capacité à reconnaître toutes les parties que forme une entité individuelle. Par exemple un *intervalle de temps* s'inscrit toujours dans une *durée*.
3. **La dépendance** : une propriété x est dépendante d'une propriété y si pour toutes les instances de x il doit nécessairement exister une instance de y, qui n'est ni une part ni un constituant de x. Ainsi *PARENT* est dépendant de *ENFANT*, car personne ne peut être parent sans avoir un enfant et vice versa.

#### 4.1.4 Les critères de **Poveda-Villalón et al. [2012]**

Plus récemment, **Poveda-Villalón et al. [2012]** (dans la continuité des travaux amorcés par **Gómez-Pérez [2001]**) ont réalisé une analyse des outils de validation disponibles et

#### 4.1 La définition des critères de validation

---

ainsi défini six dimensions/aspects qui permettent de conclure à une ontologie de qualité :

1. **La consistance logique** : elle se réfère au fait qu'il y ait (a) des inconsistances logiques, ou (b) des bouts de l'ontologie qui puissent potentiellement mener à une inconsistance, sans pour autant être détectable par un raisonneur (à moins que l'ontologie ne soit peuplée).  
Exemple : définition erronée des relations inverses, équivalentes, symétriques ou encore transitives pouvant conduire à des contradictions logiques.
2. **Les problèmes de modélisation** : se posent si l'ontologie est définie en utilisant correctement les primitives données par les langages d'implémentation d'ontologies, ou si des choix de modélisation pourraient être améliorés.  
Exemple : les classes disjointes sont-elles définies ? La granularité de l'arborescence est-elle adaptée ? Les domaines et co-domaines des propriétés sont-ils définis ?
3. **La spécification du langage ontologique** : indique si l'ontologie est conforme (par exemple que la syntaxe est correcte) avec les spécifications du langage ontologique utilisé pour implémenter l'ontologie.  
Exemple : une ontologie développée en OWL en respecte-t-elle la syntaxe ?
4. **La représentation du monde réel** : renvoie à la précision de la modélisation ontologique du domaine. Cette dimension doit être vérifiée par des humains (par exemple, les ontologues et les experts du domaine).  
Exemple : les informations modélisées sont-elles conformes à la réalité du monde ? Les ambiguïtés du langage naturel sont-elles résolues dans ce modèle ?
5. **Les applications sémantiques** : cette dimension indique si l'ontologie est adéquate pour les applications qui lui sont destinées.  
Exemple : l'ontologie contient-elle des labels pour une utilisation dans un système d'annotation ? Les règles d'inférence permettent-elle d'utiliser l'ontologie dans un système de raisonnement ? La définition des concepts au niveau logique permet-elle le partage de la compréhension de connaissances ?
6. **La compréhension humaine** : cet aspect définit si l'ontologie fournit suffisamment d'informations pour être comprise par un humain. Cet aspect est donc lié au contenu informatif (méta-donnée) de l'ontologie.  
Exemple : l'ontologie permet-elle de désambiguïser les concepts ? Les espaces de nom sont-ils unifiés ? Les noms des concepts ont-ils tous le même format ?

##### 4.1.5 Des définitions standardisées aux problèmes de qualité des ontologies, selon les travaux de Gherasim *et al.* [2012]

Gherasim *et al.* [2012] se sont penchés sur l'absence de standard dans la définition des problèmes de qualité des ontologies. Les chercheurs ont tout d'abord proposé une typologie des problèmes rencontrés lors de l'évaluation des ontologies : (1) les erreurs taxonomiques ; (2) les anomalies de conception ; (3) les anti-patterns ; (4) les Pitfalls. Cette dernière catégorie correspond aux travaux de Poveda-Villalón *et al.* [2012] que nous utilisons dans nos travaux sur la validation des ontologies, présentés au chapitre 7. Suite à cette typologie, les chercheurs ont définis un cadre pour l'identification des problèmes de

qualité dans l'ontologie, composé de deux dimensions orthogonales : (1) les erreurs (les problèmes qui compromettent l'utilisation de l'ontologie) face aux situations inadaptées (les problèmes qui ne gênent pas l'utilisation de l'ontologie) ; (2) l'aspect logique (relatif à l'aspect computationnel, structurel, « connaissance implicite » de l'ontologie) face à l'aspect social des problèmes (relatif à l'aspect sémantique, « connaissance explicite » de l'ontologie). Ce cadre regroupe 24 problèmes, divisés en 12 problèmes relatifs à l'aspect logique et 12 problèmes relatifs à l'aspect social. En annexe G, nous avons reproduit le cadre accompagné de courtes définitions des critères, ainsi que de la réponse méthodologique que nous apportons pour chaque problème (travaux présentés chapitre 7).

#### 4.1.6 Synthèse

Nous pouvons retenir de la présentation de ces quatre listes de critères que la qualité d'une ontologie se définit selon deux axes principaux :

1. La validation de la structure : correspond au respect des normes propres aux ontologies. Elle regroupe les critères de consistance logique, de problèmes de modélisation, de spécification du langage ontologique de [Poveda-Villalón et al. \[2012\]](#), ainsi que les problèmes logiques présentés dans le cadre de [Gherasim et al. \[2012\]](#) (voir annexe G).
2. La validation de la sémantique : correspond à l'adéquation de l'ontologie au domaine modélisé et au but visé. C'est-à-dire, au respect de la sémantique et de la concordance avec le but visé. Cet axe correspond aux critères de représentation du monde réel, d'applications sémantiques et de compréhension humaine de [Poveda-Villalón et al. \[2012\]](#), ainsi qu'aux problèmes de critères sociaux présentés dans le cadre de [Gherasim et al. \[2012\]](#) (voir annexe G).

## 4.2 Validation de la structure

La validation de la structure s'apparente à la validation formelle de la logique du modèle. La littérature est riche dans ce domaine et les applications développées permettent de réaliser ce travail de validation avec une intervention humaine limitée.

### 4.2.1 Les moteurs d'inférence

La validation de la structure a mené à de nombreuses études et applications. Les raisonneurs permettent de vérifier la consistance et la cohérence d'un modèle en répondant aux questions : (a) existe-t-il une interprétation qui soit un modèle de l'ontologie ainsi développée (existe-t-il un monde qui soit représenté par l'ontologie) ? ; (b) toutes les classes sont-elles satisfaites (peuvent-elles toutes être associées à des instances ?). Ainsi, les raisonneurs réalisent entre autres les trois grandes vérifications logiques suivantes :

1. **La consistance** : permet de s'assurer qu'une ontologie ne contient pas de faits contradictoires.
2. **La satisfaisabilité des concepts** : afin de déterminer s'il est possible pour une classe d'avoir des instances. Si une classe est insatisfaisable, alors définir une instance de cette classe entraînera une inconsistance de toute l'ontologie.

3. **La classification** : permet de calculer les relations de sous-classes entre chaque classe nommée, afin de créer la hiérarchie de classe complète. La hiérarchie de classe peut être utilisée pour répondre à des besoins tels que l'extraction de tout ou partie des sous-classes directes d'une classe.

L'un des outils les plus populaires et le premier à avoir pris en charge tout le OWL-DL est le raisonneur PELLET<sup>1</sup> [Sirin *et al.*, 2007]. Il a été développé pour aider à répondre aux exigences du langage ontologique OWL du W3C. Il peut s'utiliser en ligne de commande, sur le Web pour éviter toute installation, en API binding pour Jena ou en API OWL de Manchester. Nous pouvons également citer les raisonneurs HERMIT [Shearer *et al.*, 2008] et FACTPLUS [Tsarkov et Horrocks, 2006] parmi les plus utilisés en raison de leur présence en tant que *plugins* dans l'éditeur d'ontologies PROTÉGÉ<sup>2</sup>.

### 4.2.2 Les méthodes et outils

Un bon nombre de méthodes et outils ont été développés pour aider à la validation d'ontologies. Nous en présentons ici quelques unes, en particulier celles qui correspondent aux listes de critères énoncées à la section précédente.

ONTOCHECK<sup>3</sup> [Schober *et al.*, 2012] se présente comme un module d'extension à l'éditeur d'ontologies PROTÉGÉ et vise à contrôler le respect des conventions de nommage, ainsi que l'exhaustivité des méta-données. Pour ce faire, il vérifie les cardinalités, la complétude des méta-données, les labels, les conventions typographiques ainsi que les métriques de l'ontologie.

XD ANALYZER a été développé dans le cadre du projet NEON<sup>4</sup>, pour faire un retour qualitatif à l'utilisateur en suivant la méthodologie XD. Cette dernière fournit une liste de bonnes pratiques (concernant entre autres, les labels, les commentaires, les concepts non utilisés) à respecter pour la construction d'ontologies. RADON [Ji *et al.*, 2009] a également été ajouté en module d'extension à NEON TOOLKIT afin de faciliter les tâches de vérification de la consistance.

ONTOCLEAN [Guarino et Welty, 2000] est une méthodologie permettant la validation de l'adéquation des relations taxonomiques d'une ontologie. La première étape consiste à annoter les concepts selon les méta-propriétés de rigidité, d'unité et de dépendance (présenté en section G). Ensuite, une analyse basée sur des contraintes prédéfinies est réalisée sur les annotations afin de mettre en avant les erreurs taxonomiques. ONTOCLEAN a été implémentée dans deux principales applications : ODECLEAN et AEON. Guarino et Welty [2009] note que le but d'ONTOCLEAN est d'amener les ontologues à penser les conséquences logiques de certains choix de modélisation.

---

1. <http://clarkparsia.com/pellet/>  
2. <http://protege.stanford.edu/products.php>  
3. <http://protegewiki.stanford.edu/wiki/OntoCheck>  
4. <http://neon-toolkit.org/wiki/XDTools>

ODECLEAN [Fernández-López et Gómez-Pérez, 2002] se présente comme un module disponible sous l'éditeur d'ontologies WebODE. Premièrement, les auteurs ont développé une top ontologie des universaux en suivant la méthode ONTOCLEAN. Ils y ont donc incluse les méta-propriétés. Deuxièmement, ils ont ajouté à la top ontologie des universaux les règles de contraintes associées aux méta-propriétés d'ONTOCLEAN, via leur éditeur d'axiomes et de règles WAB. Ils ont ainsi construit une ontologie qu'ils ont ensuite convertie en PROLOG via WAB. L'utilisateur de WEBODE peut choisir d'utiliser les principes d'ONTOCLEAN et annoter chaque concept avec sa valeur en tant que méta-propriété. L'ontologie peut ensuite être validée automatiquement grâce aux annotations réalisées par l'utilisateur et au module ODECLEAN qui implémente les contraintes associées aux méta-propriétés.

AEON [Völker *et al.*, 2008] met en avant les contraintes de temps liées à la méthode ONTOCLEAN. En effet, la première étape oblige à l'annotation manuelle des concepts et à une intervention d'ontologues particulièrement expérimentés. Le but d'AEON est donc d'annoter automatiquement les concepts de l'ontologie en suivant la méthode ONTOCLEAN, puis de réaliser la vérification des contraintes. Les auteurs réalisent l'annotation automatique des concepts, par le biais d'une concordance lexico-syntaxique réalisée sur des documents disponibles sur le Web. Le but est d'obtenir des informations sur la rigidité, l'unité, la dépendance et l'identité d'un concept de l'ontologie. Les auteurs considèrent que le Web peut être utilisé comme un corpus en langage naturel apte à résoudre les questions autour de l'essence, l'unité et l'identité d'un terme tel que définie dans ONTOCLEAN. Cependant, cette méthode s'est trouvée confrontée aux problèmes liés à l'ambiguïté du langage (par exemple, le mot « verre » qui peut être soit un « verre d'eau » soit un « verre de lunette » et qui selon le premier sens est un tout (unité) et selon le second ne l'est pas (anti-unité)).

LE PROJET PERTO MED avait pour objectif la construction d'une ontologie de la pneumonie [Baneyx, 2007]. Dans le cadre de ce projet, Baneyx et Charlet [2006] se sont intéressés aux problématiques liées à l'évaluation d'une ontologie. Ils ont retenu les critères de Gómez-Pérez [2004] afin de mettre en lumière la cohérence, la complétude et la précision de leur modèle. Les auteurs définissent la validation comme le moyen de s'assurer que l'ontologie modélise, non pas le monde réel, mais un modèle des connaissances que nous avons sur le monde. La méthode adoptée est purement manuelle et se fonde sur les connaissances des experts (cohérence), l'alignement de la taxonomie avec un thésaurus spécialisé (complétude), l'adéquation de l'ontologie à son SBC (précision).

ONTOLOGY PITFALL SCANNER! (OOPS!) <sup>5</sup> [Poveda-Villalón *et al.*, 2012] est un outil indépendant de tout éditeur d'ontologies. Il s'utilise uniquement en ligne. Le but de OOPS! est l'identification des anomalies ou mauvaises pratiques dans une ontologie. Pour cela, les auteurs ont défini un certain nombre de « pitfalls » (embûches, pièges) répertoriés en langage naturel dans un catalogue <sup>6</sup>. Nous en comptons actuellement 40, dont 32 sont implémentés en tant que classes java ajoutées au module d'analyse des pitfalls. En entrée,

5. <http://oops.linkeddata.es/>

6. <http://oops.linkeddata.es/catalogue.jsp>

l'application prend l'URI d'une ontologie ou bien le code source en RDF<sup>7</sup>. L'ontologie est chargée via l'API Jena avant d'être analysée pour en extraire les erreurs potentielles. Le résultat est une page répertoriant les erreurs selon le pitfall identifié, avec une proposition de résolution de l'erreur. Les pitfalls peuvent concerner des éléments individuels, plusieurs éléments ou toute l'ontologie. Une méthodologie du même type avait déjà été utilisée avec succès, pour valider une ontologie développée au sein de notre équipe [Charlet *et al.*, 2012].

## 4.3 Validation de la sémantique

La validation de la sémantique est un sujet nettement moins riche dans la littérature. Elle s'apparente à la validation de la sémantique des connaissances, la validation du modèle conceptuel en adéquation avec la réalité qu'il modélise. Cette étape oblige à une communication accrue entre acteurs de domaines d'expertises différents : les ontologues d'un côté et les spécialistes du domaine modélisé dans l'ontologie de l'autre. Alors que la validation structurelle est aisément réalisée à l'aide de méthodes automatiques, qui ne demandent quasiment aucune intervention humaine ; les méthodes de validation sémantique obligent à trouver des stratagèmes, afin d'alléger l'implication des ontologues et surtout des experts.

### 4.3.1 Un développement collaboratif couplé à des approches ergonomes en réponse aux difficultés liées à la validation sémantique

Ghidini *et al.* [2012] précisent que dans le cas de la description de modèles d'entreprises, une seule personne ne peut posséder toutes les connaissances et compétences lui permettant de modéliser l'entière du domaine. Les auteurs se sont alors fondés sur un système de wiki<sup>8</sup>, pour développer un outil de construction collaborative d'ontologies. Chaque élément du modèle est décrit dans une page MOKI au travers d'informations structurées, qui peuvent être comprises par n'importe quel acteur ayant des connaissances techniques ou non. En effet, chaque page contient une description informelle de l'élément sous forme de texte libre, d'images ou de dessins, et une partie structurée (par le biais d'un formulaire) dans laquelle les éléments sont décrits sous forme de triplets (sujet, relation, objet). De plus, l'outil intègre un système de validation sémantique des inférences logiques par questionnaires [Pammer *et al.*, 2010].

Le développement collaboratif a également été adopté par Ressad-Bouidghaghen *et al.* [2013], qui explique qu'il est plus facile de travailler de façon collaborative, lors de grands projets de construction de ressources sémantiques. En effet, cela permet d'établir dès la construction, une « modélisation consensuelle » acceptée et validée par les différents acteurs. Ils mettent également en avant la construction modulaire, afin que chaque acteur puisse intervenir dans son domaine ou sous-domaine de compétence. Pour chaque module, un ontologue est désigné comme acteur responsable. Lors de l'intégration des modules, les choix qui font débat ou qui se heurtent à la modélisation d'autres modules sont

---

7. À noter qu'il est possible de déclarer qu'on ne souhaite pas que le code chargé sur la page soit conservé.

8. Un wiki est un site Web dynamique dont la création, l'édition, la modification des pages peuvent être effectuées par les utilisateurs du site.



discutés et votés. Le but est d'établir un consensus autour du point de vue adopté pour la modélisation. Cette méthode permet de garder une trace des décisions, de développer plusieurs modules en parallèle, et d'établir un consensus rapide autour des questions de sémantique, de lexique ou de modélisation.

#### 4.3.2 Les outils collaboratifs pour la validation de la sémantique

Concernant les outils pour le développement collaboratif d'ontologies, nous pouvons citer :

- le serveur ONTOLINGUA<sup>9</sup> : propose un environnement collaboratif pour parcourir, créer, éditer, modifier ou utiliser les ontologies [Farquhar *et al.*, 1997]
- WEBPROTÉGÉ<sup>10</sup> : a été développé en reprenant l'architecture de Protégé. Il permet également le développement collaboratif d'ontologies et est accessible via n'importe quel navigateur web [Tudorache *et al.*, 2013]
- Ressad-Bouidghaghen *et al.* [2013] : indique qu'un module est en cours de construction pour permettre le développement collaboratif avec TERMINAE<sup>11</sup>.

#### 4.3.3 Les limites du « tout collaboratif » pour la validation de la sémantique

Cependant, le développement collaboratif a ses limites. Dans le cas des méthodes de construction entièrement collaboratives, il demande une grande disponibilité des acteurs. Et dans le cas de la construction d'ontologies, il implique de posséder un certain nombre de compétences techniques non négligeables, notamment en formalisme logique. C'est le constat que fait Ben Abacha *et al.* [2013] qui arguent que la collaboration avec des médecins oblige à trouver des stratagèmes, leur permettant de s'impliquer dans la validation collaborative de l'ontologie, sans qu'ils n'aient à toucher au modèle. Sur ce constat, ils proposent donc un système de validation par questions/réponses, et invalidation par texte libre. Le médecin se trouve face à une liste de questions booléennes. Lorsqu'il invalide une question, il peut fournir une justification sous forme de texte libre, afin de permettre à l'ontologue de corriger le formalisme incriminé. À noter que les auteurs partent du postulat qu'une réduction de la communication entre experts et ontologues réduit les erreurs. En l'état actuel, leur méthode se trouve limitée par le grand nombre de questions générées pour des ontologies de taille de plusieurs milliers de concepts. Pour pallier ce problème, les auteurs proposent de mettre en place une validation au fur et à mesure de la construction de l'ontologie.

### 4.4 Synthèse

Dans ce chapitre, nous nous sommes intéressés à la validation des ontologies : ce qu'elle signifie et la manière dont elle se concrétise.

Nous retenons que dans la littérature, la qualité d'une ontologie se définit par la validation de sa structure (qui correspond au respect des normes propres aux ontologies) et

9. <http://www.ksl.stanford.edu/software/ontolingua/>

10. <http://webprotege.stanford.edu/>

11. <http://lipn.fr/terminae/index.php/Download>

la validation de sa sémantique (qui correspond à l'adéquation de l'ontologie au domaine modélisé et au but visé). À cela s'ajoute des listes exhaustives de critères de validation qui permettent de s'accorder sur la notion d'ontologie de qualité, en particulier au travers des travaux plus récents tels que ceux de [Poveda-Villalón et al. \[2012\]](#) et [Gherasim et al. \[2012\]](#). Par ailleurs, nous avons constaté que toutes ces méthodes pouvaient être divisées en méthodes pour valider la structure [[Guarino et Welty, 2000](#), [Fernández-López et Gómez-Pérez, 2002](#), [Völker et al., 2008](#), [Shearer et al., 2008](#), [Schober et al., 2012](#), [Poveda-Villalón et al., 2012](#)] et méthodes pour valider la sémantique [[Pammer et al., 2010](#), [Ghidini et al., 2012](#), [Ressad-Bouidghaghen et al., 2013](#), [Ben Abacha et al., 2013](#)].

Alors que la validation de la logique, de la structure des ontologies peut être aisément réalisée à l'aide de méthodes semi-automatiques, les méthodes actuelles pour la validation de la sémantique des ontologies laissent de nombreuses questions en suspens. Le développement collaboratif semble être la meilleure des solutions proposées, mais elle n'est humainement pas toujours envisageable. La proposition de [Ben Abacha et al. \[2013\]](#) de passer par une interface graphique pour faciliter la collaboration avec les acteurs du domaine semble être une stratégie pertinente. Néanmoins, il n'est pas démontré qu'elle est applicable à de larges modèles de connaissances. En outre, la validation des axiomes logiques est peu discutée dans la littérature. Dans nos travaux, nous développons des ontologies légères, qui contiennent un grand nombre de classes avec une structure complexe, de nombreuses définitions et de nombreux labels. Nos ontologies n'ont pas vocation à faire du raisonnement, elles contiennent donc peu d'axiomes logiques. Nous avons ainsi limité notre analyse du domaine sur la validation de la sémantique, à la validation des ontologies dites « légères » (voir section 2.2.2). En outre, nous constatons au travers de notre état de l'art que les outils qui permettent de valider la sémantique de l'ontologie font toutes appel à des personnes spécialistes, dans le domaine modélisé par l'ontologie à valider. À l'heure actuelle, aucune méthode ne peut se passer de l'humain, pour valider l'adéquation d'une ontologie avec la réalité qu'elle tend à modéliser.

Ainsi, aucune méthode ne permet à ce jour de vérifier la qualité de l'ontologie sous tous ses aspects, structurels et sémantiques, logiques et sociaux. Dans le chapitre 7, nous proposons un cadre méthodologique pour vérifier la qualité d'une ontologie, c'est-à-dire pour vérifier que l'ontologie répond aux critères de qualité posés dans la littérature. Nous avons réalisé un chaînage de plusieurs méthodes de validation, que nous avons sélectionné selon les critères qu'elles valident. Nous constatons également que les méthodes ne proposent pas de support pour travailler avec les experts, à la validation du modèle de connaissances. Nous avons ajouté une étape à notre cadre méthodologique, pour décrire des entretiens semi-directifs réalisés avec des experts.





## **Deuxième partie**

# **Contributions scientifiques théoriques et pratiques**



# Chapitre 5

## Construction du module ontologique « facteurs sociaux et environnementaux des maladies psychiatriques » (OntoPsychiaFSE)

### Sommaire

<b>5.1 Choix de ce module</b>	<b>100</b>
5.1.1 Enjeux	100
5.1.2 Objectifs	101
<b>5.2 Présentation du corpus</b>	<b>101</b>
5.2.1 Composition du corpus	101
5.2.2 Le choix de ce corpus	102
<b>5.3 Méthode de construction par approche hybride</b>	<b>104</b>
5.3.1 Extraction de termes candidats avec la méthode TERMINAE	105
5.3.2 Conceptualisation du domaine	107
5.3.3 Enrichissement du modèle	107
<b>5.4 Résultats</b>	<b>108</b>
5.4.1 Extraction de termes candidats avec la méthode TERMINAE	108
5.4.2 Conceptualisation du domaine	109
5.4.3 Enrichissement du domaine	113
<b>5.5 Synthèse</b>	<b>113</b>

*La première partie de ce manuscrit, nous a permis de présenter notre projet en contexte des recherches et avancées récentes réalisées dans les domaines connexes à notre étude : l'organisation des connaissances au travers des ontologies, ainsi que les méthodes de construction et de validation d'ontologies. Dans cette deuxième partie, nous abordons le cœur du travail théorique et pratique réalisé dans cette thèse à partir de l'état de l'art. Nous avons choisi en tout premier lieu, de réaliser ONTOPSYCHIA sous forme de modules : (1) facteurs sociaux et environnementaux et (2) troubles/maladies mentales. Ce choix se justifie par toutes les raisons évoquées au Chapitre 3, telles que faciliter le développement, la réutilisation et la maintenance de l'ontologie. En outre, nous avons constaté au cours de notre étude de la littérature (voir l'introduction et la section 1.3) que les thésaurus et les ontologies médicales, qui modélisent des aspects sociaux dans la vie d'un patient restent encore trop peu exploités en psychiatrie. Pourtant, le développement de la médecine de précision met en avant l'importance de tendre vers une approche holistique, dans la prise en charge des patients en psychiatrie. Ainsi, l'analyse de la prévalence et de l'incidence des facteurs de risque sociaux et environnementaux des maladies psychiatriques est cruciale pour les comprendre et les traiter et peut avoir des impacts significatifs sur les décisions politiques (coûts et durées des hospitalisations notamment). Dès lors, comment pouvons nous modéliser ces facteurs sociaux ? Quelle méthodologie adopter en l'absence de modèle déjà existant ? Dans ce chapitre, nous présentons les différentes étapes de la conceptualisation du module sur les facteurs sociaux et environnementaux (OntoPsychiaFSE), à partir de notre corpus composé de comptes rendus d'hospitalisation issus de l'hôpital Sainte-Anne. Nous suivons les étapes de la méthodologie hybride TOREUSE2ONTO [Drame, 2014] présentée en section 3.4.2. Nous avons suivi les engagements ontologiques de la méthode ARCHONTE (présentés en sections 2.2.2 et 3.2.2) [Bachimont et al., 2002], et en particulier le principe de normalisation sémantique, pour développer la conceptualisation du domaine. La première section de ce chapitre aborde les enjeux et objectifs liés au développement de ce module. La deuxième section présente le matériel utilisé pour cette étude. La troisième section décrit pas à pas la méthode utilisée pour le développement. Enfin, une quatrième section présente les résultats obtenus lors de l'application de la méthode pour la construction du module.*

## 5.1 Choix de ce module

### 5.1.1 Enjeux

La médecine de précision permet de définir des sous groupes dans la population à l'aide de marqueurs biologiques ou génétiques [Picard, 2014]. Cette stratification de la médecine est rendue possible grâce aux développements des nouvelles technologies, ainsi qu'aux progrès réalisés en génétique, neurosciences, cognition et santé mentale. Elle a pour but de soigner les patients en fonction des caractéristiques de leur pathologie et des spécificités génétiques et environnementales, qui peuvent différer d'un individu à l'autre. Bourgin et Duchesnay [2014] se sont intéressés aux éléments qui permettent d'identifier des sujets à haut risque de psychose, afin de prévenir une prise en charge précoce, et ainsi réduire les risques de transition psychotique. Dans le cadre de cette étude, les auteurs ont souligné l'enjeu que représente la médecine de précision, pour la psychiatrie. Elle permet en effet de tendre vers une approche holistique, dans la prise en charge des patients en psychiatrie.

### 5.1.2 Objectifs

Lors de la première lecture des comptes rendus d'hospitalisation (CRH), nous avons constaté qu'une part non négligeable du compte rendu était destiné à la narration d'éléments biographiques, concernant le patient et sa famille. La situation sociale, professionnelle ou encore les liens que le patient entretient avec son entourage sont décrits, pour mettre en lumière le contexte social dans lequel évolue le patient. Nous avons donc fait le choix du module sur les facteurs sociaux et environnementaux, afin de définir un cadre pour l'étude d'une possible corrélation entre les événements de la vie sociale et les troubles de la santé mentale. Dans cette ontologie, nous décrivons des facteurs sociaux ou environnementaux qui peuvent ou non affecter la vie d'un patient, de manière positive, négative ou variable. Le but à long terme est de disposer d'un outil, pour l'analyse de ces facteurs, afin d'aider à la prévision ou à l'identification des troubles ainsi qu'à la prise en charge des patients.

## 5.2 Présentation du corpus

Dans cette section, nous présentons notre corpus, l'organisation des connaissances au sein de ce corpus et les raisons du choix de ce corpus. Une anonymisation des documents a été réalisée au préalable à notre étude. Elle est présentée en annexe [H](#).

### 5.2.1 Composition du corpus

Le corpus qui sert de base à nos travaux est composé d'un peu plus de 8 700 comptes-rendus d'hospitalisation (CRH), issus du service hospitalo-universitaire de santé mentale et thérapeutique de l'hôpital Sainte-Anne. Ces CRH sont pseudo-standardisés et sous format Word. Ils couvrent une période de dix années. Le diagnostic est donné en fin de CRH depuis quelques années, ainsi que le codage selon la CIM-10. Parallèlement, pour chaque patient, l'activité est recueillie par le département d'Informatique Médicale (DIM) selon le codage EDGAR (Entretien, Démarche, Groupe, Accompagnement, Réunion) [[Richard et al., 2013](#)].

Un CRH est organisé en différentes parties. Nous remarquons qu'une part non négligeable de l'information contenue dans ces CRH est laissée de côté et ignorée dans les méthodes de codage actuelles, en particulier la comorbidité. En effet le diagnostic se conclut par un codage CIM-10 et ne tient pas compte de toutes les observations qui ont pu être faites durant l'hospitalisation du patient.

1. En tête : contient les informations relatives à l'hôpital, l'identité du patient, sa date de naissance, les dates d'entrée et de sortie du service.
2. Le corps du CRH :
  - **Modalité d'hospitalisation** : indique si le patient a été hospitalisé volontairement ou non (Hospitalisation Libre -HL-, Hospitalisation à la Demande d'un Tiers -HDT-, Hospitalisation d'Office -HO-).

- **Motif d'hospitalisation** : indique la ou les raisons qui ont entraînées cette hospitalisation. Exemple (extrait) : « prise en charge pour ingestion médicamenteuse » ; « recrudescence délirante » ; « rééquilibration du traitement » ; etc.
- **Biographie** : retrace les éléments biographiques généraux du patient, tels que le contexte familial, scolaire, professionnel, culturel.
- **Antécédents médico-chirurgicaux personnels** : retrace tous les antécédents médico-chirurgicaux du patient. Exemple : « fracture de la clavicule gauche » ; « Tabagisme actif » ; « Allergie possible au antalvic ».
- **Antécédents psychiatriques personnels** : retrace tous les antécédents psychiatriques du patient. Exemple : « hospitalisation en HDT de X à l'hôpital X pendant X jours » ; « Trouble bipolaire depuis X » ; « X : Début des troubles ». « X » correspond à un terme anonymisé.
- **Antécédents familiaux** : apporte des éléments médicaux sur l'entourage familial du patient. Exemple : « Frère a fait une tentative de suicide » ; « Cousine décédée par suicide ».
- **Histoire de la maladie** : retrace les évolutions de la maladie ou des maladies. Exemple : « Sentiment de tristesse très fluctuant avec idéations suicidaires et une note d'agressivité verbale ».
- **Examen à l'entrée** : dresse l'état psychologique général du patient à l'entrée dans le service. Exemple : « Exaltation désordonnée. »
- **Évolution dans le service** : retrace le parcours du patient durant son hospitalisation. Exemple : « Espacement progressif des ECT à partir du mois d'août, amélioration progressive de la symptomatologie anxio dissociative avec disparition des idées suicidaires »
- **Examens complémentaires** : liste des résultats des examens complémentaires réalisés durant l'hospitalisation.
- **Inclusion dans un protocole** : oui non (case à cocher).
- **Au total** : dresse le bilan de l'hospitalisation du patient.
- **Traitement de sortie** : indique le traitement mis en place à la sortie de l'hôpital.
- **Diagnostic** : codage CIM-10.

### 5.2.2 Le choix de ce corpus

Dans cette section, nous présentons les points forts de notre corpus. Nous précisons également les limites qu'il nous impose et les moyens de les dépasser.

#### Les « précautions » d'usage :

« La constitution d'un corpus est très délicate de manière générale, car le corpus conditionne largement le type et la nature des traitements que l'on peut effectuer sans que l'on ait forcément loisir de choisir le type de données le plus adéquat. Le choix d'un corpus introduit des biais sans qu'il soit toujours loisible de les apprécier. ». Cette citation de Bachimont [2000] est reprise par Aimé [2015] pour appuyer « la difficulté de choisir

## 5.2 Présentation du corpus

---

un corpus ». Ce dernier énumère également les « précautions méthodologiques » décrites dans Bourigault *et al.* [2004], afin de parvenir à construire une ressource documentaire la plus adéquate au projet :

**La consensualité** : vise à assurer que les documents sélectionnés font consensus au sein de la communauté spécialistes ou experts de ce domaine. Et ce, dans le but de se protéger d'une quelconque remise en cause du projet pour motif de corpus « inadapté ».

**La taille** : vise à s'assurer que le corpus couvre l'ensemble du domaine. Avec le développement des outils du traitement automatique du langage (TAL) pour traiter automatiquement les corpus, la taille de ceux-ci tend à croître. Bourigault *et al.* [2004] préconise d'opter pour un corpus « suffisamment petit ou redondant pour pouvoir être appréhendé de façon globale par l'analyste », avec un nombre de mots compris entre 50 000 et 200 000.

**Les critères de disponibilité** : [Aimé, 2015] ajoute à cela les différents critères qui permettent d'affirmer la disponibilité d'un corpus : (1) « critère de connaissance » qui est attesté par les experts ; (2) « critère légal » lié à l'usage du corpus dans un but de recherche ; (3) « critère social » lié au contexte de création du document.

### L'adéquation de notre corpus :

Notre ontologie porte sur le domaine de la psychiatrie, nous avons donc besoin d'un corpus représentatif de ce domaine, son vocabulaire et ses notions, tout en respectant les contraintes énumérées précédemment.

**La consensualité de notre corpus** est attestée par sa nature. En effet, les comptes-rendus d'hospitalisation qui composent notre corpus sont rédigés par les praticiens en charge des patients ; et ils sont destinés au suivi des patients au sein de la structure hospitalière. Ils se doivent donc d'être exhaustifs, détaillés, et fournir un maximum d'informations sur la/les maladies. Ces documents sont rédigés par des spécialistes du domaine et ils s'adressent à l'ensemble du personnel soignant. Nous sommes donc assuré de l'utilisation d'un vocabulaire spécialisé (y compris sigle et abréviation du domaine). Enfin, les CRH s'adressent également à la famille du patient et au patient lui-même. Ils doivent donc être suffisamment vulgarisés, pour être compréhensible par une personne extérieure au domaine.

**La taille de notre corpus** est d'environ 8 700 CRH. Nous couvrons largement notre domaine d'étude en termes de données. Cela peut même sembler trop important, malgré l'automatisation des tâches d'extraction des connaissances. La totalité des CRH compte près de 5,8 millions de mots (5 788 637), dont 73 217 mots différents. Ces chiffres ont été calculés sur les données brutes et ne tiennent pas compte du bruit : mots mal orthographiés, erreurs de ponctuation ou encore données numériques. Ils permettent de communiquer un ordre de grandeur du corpus.



**La disponibilité de notre corpus**, selon le critère de connaissance, a été confirmé à la phase de validation réalisée avec l'équipe du SHU. Cette validation a mis en lumière l'adéquation des connaissances issues du corpus, avec les attentes des praticiens. Le critère légal est respecté par l'autorisation de la CNIL et le travail d'anonymisation des CRH réalisé au début du projet (ce travail est décrit en annexe ).

#### **Les limites de notre corpus :**

Notre corpus présente néanmoins un certain nombre de limites.

**La consensualité de notre corpus** est acceptable dans l'enceinte de l'hôpital Sainte-Anne, mais il ne saurait en être de même au sein de tous les hôpitaux psychiatriques. En effet, la psychiatrie est un vaste domaine régi par des écoles et courants de pensée différents. Les deux principaux sont les courants biologique et psychologique. [Aimé \[2015\]](#) cite également les courants cognitiviste, béhavioriste ou encore humaniste, et insiste sur le décalage auquel font face certains courants. Dès lors, il devient compliqué d'aligner les références conceptuelles et d'établir un consensus. Nous pouvons préciser que notre corpus, tout comme notre étude, se place dans le courant de pensée psychiatrique du Service Hospitalo Universitaire de l'hôpital Sainte-Anne, qui est celui de la biologie.

**La taille importante de notre corpus** a limité l'automatisation des traitements. En effet, certains outils n'ont pu supporter ce corpus. Nous souhaitions travailler sous l'environnement TERMINAE au prétexte de ce projet. Toutefois, après des mois de tentatives infructueuses, nous avons été contraint d'abandonner l'idée, pour nous tourner vers des méthodes plus robustes à la taille du corpus. L'utilisation d'outils TAL, pour l'extraction automatique de concepts doit se penser en adéquation au corpus. Ce dernier est d'une taille conséquente, plus de 5 millions de mots. Les traitements nécessaires à l'extraction d'une liste de concepts candidats sont lourds, qu'ils soient automatiques ou manuels. Ils nécessitent donc des ressources matérielles et logicielles adaptées.

**La disponibilité de notre corpus** a été une étape déterminante en termes de critères légaux. L'obtention des autorisations légales quand il s'agit d'utiliser des données médicales peut se trouver être une étape délicate, à laquelle il est nécessaire de consacrer un temps conséquent et parfois décisif à l'échelle du projet. En pratique, l'utilisation du corpus n'était possible que si nous respections les règles de la Commission Nationale Informatique et Libertés (CNIL), en particulier le fait que le corpus soit anonymisé. Cette problématique fait l'objet de la section [H.1.1](#).

### **5.3 Méthode de construction par approche hybride**

Pour développer notre ontologie, nous avons suivi les grandes étapes de la méthodologie hybride TOREUSE2ONTO [[Drame, 2014](#)] présentée en section [3.4.2](#), qui s'inspire notamment de la méthode ARCHONTE [[Bachimont et al., 2002](#), [Charlet et al., 2006](#)]. Nous avons suivi cette dernière méthode pour le développement de la conceptualisation du domaine, car elle en détail toutes les étapes (voir en section [3.2.2](#)). Nous avons décidé de

mettre de côté les classifications au cours de la construction de l'ontologie, pour nous concentrer sur l'information contenue dans notre corpus (approche descendante), et revenir aux classifications de domaine uniquement pour la phase d'enrichissement (approche ascendante).

#### 5.3.1 Extraction de termes candidats avec la méthode TERMINAE

Nos travaux intègrent la construction d'une ressource termino-ontologique (RTO). Les méthodes d'extractions fondées sur des thésaurus ne répondent pas à nos besoins, car ils dépendent des classifications de référence. Nous avons donc opté pour une méthode par extraction de termes candidats, pour rester indépendant des terminologies en usage.

Notre choix s'est porté en premier lieu sur la méthode TERMINAE (présentée en section 3.2.2) et l'extracteur de termes candidats (ETC) YATEA (présenté en section 3.2.3). Ces outils, utilisés conjointement, permettent de construire une RTO et le modèle conceptuel associé. Nous avons opté pour l'ETC YATEA, car il était important pour nous de pouvoir installer le logiciel sur une machine que nous maîtrisons, à l'opposé d'un traitement sur serveur, afin de garantir l'anonymat des données de notre corpus. L'outil YATEA requiert en entrée, un corpus segmenté en mots et accompagnés d'une annotation morpho-syntaxique réalisée par l'outil TREETAGGER. Il fournit en sortie, plusieurs fichiers au format XML, texte et HTML. L'annotateur TREETAGGER donnait de mauvais résultats sur le corpus d'étude en français : de nombreux mots n'étaient pas reconnus et annotés comme tel. La suite des traitements jusqu'à l'extraction des termes candidats était donc rendue difficile, voire nulle en termes de résultat.

Nous avons donc opté pour l'annotateur MELTMELT<sup>1</sup> [Denis et Sagot, 2012] qui réalise l'annotation morphosyntaxique de corpus en français. Ce logiciel a été développé au laboratoire d'Analyse linguistique profonde à grande échelle<sup>2</sup> (Alpage) par Pascal Denis et Benoît Sagot. MELT s'appuie « sur un modèle probabiliste séquentiel qui bénéficie d'informations issues d'un lexique exogène, à savoir le Lefff » [Denis et Sagot, 2010]. Le Lefff est un lexique morphologique et syntaxique. C'est-à-dire qu'à chaque lemme sont associées des informations concernant la morphologie du mot et la syntaxe. [Denis et Sagot, 2012] qui offre un meilleur taux d'annotation sur notre corpus spécialisé. Nous avons ensuite développé un convertisseur de formats et d'étiquettes, afin que le corpus annoté par MELT soit utilisable sous YATEA.

Les outils, que nous utilisons pour la suite du traitement du corpus, acceptent uniquement les étiquettes de l'annotateur TREETAGGER. Nous avons converti les étiquettes de MELT en étiquettes de TREETAGGER, afin de pouvoir procéder à l'extraction des termes candidats. Les difficultés d'annotation des CRH sont dues principalement à la simplification de la syntaxe écrite. Le texte prend parfois la forme de notes ou de listes de mots-clés, sans syntaxe pour les lier.

Afin d'illustrer les différences de traitement entre les deux annotateurs, nous avons annotés la même phrase « Trouble de l'adaptation avec humeur anxiodépressive chez un patient de 40 ans, souffrant d'un trouble de la personnalité. », avec l'annotateur MELT et avec l'annotateur TREETAGGER.

---

1. [https://gforge.inria.fr/frs/?group\\_id=481](https://gforge.inria.fr/frs/?group_id=481)

2. <http://www.inria.fr/equipes/alpage>

TABEAU 5.1 – Comparaison des résultats des annotateurs morphosyntaxiques TREETAGGER et MELT.

Mot	TREETAGGER		MELT	
	Étiquette	Lemme	Étiquette	Lemme
Trouble	NOM	trouble	ADJ	trouble
de	PRP	de	P	de
l'			DET	le
l'adaptation	NOM	<unknown>		
adaptation			NC	adaptation
avec	PRP	avec	P	avec
humeur	NOM	humeur	NC	humeur
anxiodépressive	ADJ	<unknown>	ADJ	*anxiodépressif
chez	PRP	chez	P	chez
un	DET :ART	un	DET	un
patient	ADJ	patient	NC	patient
de	PRP	de	P	de
40	NUM	@card@	DET	*40
ans	NOM	an	NC	an
,	PUN	,	PONCT	,
souffrant	VER :ppre	souffrir	VPR	souffrir
d'un	NOM	<unknown>		
d'			P	de
un			DET	un
trouble	ADJ	trouble	NC	trouble
de	PRP	de	P	de
la	DET :ART	le	DET	le
personnalité	NOM	personnalité	NC	personnalité
.	SENT	.	PONCT	.

Dans le tableau 5.1, nous constatons que TREETAGGER sort cinq erreurs d'analyse et d'annotation : « l'adaptation », « anxiodépressive », « patient », « d'un », « trouble ». MELT ne fait qu'une erreur sur la première occurrence de « Trouble » qui commence la phrase sans déterminant et qui est un cas d'ambiguïté. Cette analyse réalisée sur une phrase courte permet de mettre en avant les difficultés d'annotation avec TREETAGGER, sur un corpus de grande envergure tel que le notre, constitué pour rappel d'environ 8 700 CRH. Nous avons donc choisi de travailler avec l'annotateur morphosyntaxique MELT. Pour cela, nous avons établi une table de correspondance entre les étiquettes de TREETAGGER et celles de MELT (visible en annexe I). La conversion est ensuite réalisée automatiquement à l'aide d'un script écrit en langage de programmation Perl.

#### 5.3.2 Conceptualisation du domaine

Afin de conceptualiser le domaine en partant de la liste de termes extraites des CRH, nous avons suivi les recommandations de ARCHONTE [Bachimont *et al.*, 2002]. Cette méthode propose une assistance en quatre étapes, pour le développement de la conceptualisation d'une ontologie (voir en section 3.2.2). Tout d'abord, nous avons construit un glossaire des termes validés en concepts, associés à leur définition dans le langage courant, et toutes les équivalences lexicales, tel que les synonymes et les acronymes. Ces termes deviennent les concepts de l'ontologie et chacun d'eux est dénoté par un label préférentiel unique en anglais et un label préférentiel unique en français, et un ou des labels alternatifs (synonyme, acronyme). Ensuite, nous avons construit les relations de subsomption, soit la taxonomie de concepts. Enfin, nous avons construit les relations entre ces concepts. À cette étape, nous ne nous sommes pas occupés de la description des classes, afin de ne pas réduire l'interprétation du domaine, avant la validation des concepts et de la taxonomie par les experts.

#### 5.3.3 Enrichissement du modèle

Selon la méthode TOREUSE2ONTO, le dernière étape de la conceptualisation (avant la validation) est l'enrichissement du modèle. Cet enrichissement peut être réalisé par alignement sur les concepts existants de l'ontologie, ou par intégration de nouveaux concepts à partir d'autres ressources.

Dans le cadre de son projet de thèse, Maaroufi [2016] a étudié les techniques d'alignement de données. L'auteure rappelle que l'alignement « consiste en la recherche de correspondances entre les différents éléments de deux ensembles hétérogènes ou plus ». Le but est donc de mettre en évidence les équivalences sémantiques effectives ou inférées entre deux ensembles de données. Maaroufi [2016] cite différents types de correspondance, telles que les correspondances exactes, partielles, conditionnées ou éclatées. Ainsi que différentes techniques utilisées pour réaliser l'alignement : linguistiques, structurelles, basée sur les instances. L'auteure indique que ce sont les techniques de mesure de similarité entre les chaînes de caractères (technique linguistique), qui sont les plus plébiscitées, pour l'alignement de données. À ce titre, l'outil ONAGUI<sup>3</sup> (Ontology Alignment Graphical User Interface) est tout a fait adapté pour réaliser un alignement entre deux ontologies formalisées en SKOS (présenté en section 2.2.3) ou OWL (présenté en

---

3. <https://github.com/lmazuel/onagui> <https://sourceforge.net/projects/onagui/>

section 2.2.2). Ce logiciel open source a été développé en 2009 par Laurent Mazuel et Jean Charlet [Mazuel et Charlet, 2009]. ONAGUI utilise trois algorithmes de mesure de similarité tels qu'un alignement exact, la I-Sub distance et la distance de Levenshtein<sup>4</sup> qui a par ailleurs été utilisée avec succès dans le projet de Maaroufi [2016]. En sortie, l'outil propose une liste d'alignements entre les concepts des deux ontologies, exportable aux formats RDF, SKOS et CSV.

## 5.4 Résultats

### 5.4.1 Extraction de termes candidats avec la méthode TERMINAE

Actuellement, le logiciel YATEA ne fournit aucune interface aidant à la validation des termes. Nous avons également eu beaucoup de peine à analyser notre corpus avec ce logiciel, car il n'en a pas supporté la taille (d'environ 38,2 Mo en texte brut et d'environ 98 Mo en texte annoté morpho-syntaxiquement) et le manuel d'utilisation du logiciel n'indiquait pas de limitation de taille de corpus.

Nous avons dû découper le corpus en différentes sections thématiques, pour l'analyser. Nous avons ensuite regroupé ces résultats dans un même fichier pour l'ouvrir sous TERMINAE.

La plateforme TERMINAE permet de visualiser et analyser les résultats de l'extraction des termes candidats de YATEA. Elle présente les termes candidats en donnant des informations, sur le nombre d'occurrences et le contexte d'apparition. Ce contexte peut s'avérer très utile pour désambiguïser les termes et définir des relations. Cependant, nous n'avons jamais pu ouvrir la totalité de notre corpus via cette plateforme. Nous avons discuté régulièrement avec la personne en charge du développement de l'outil, mais nous n'avons pu trouver de solution technique pour résoudre ce problème. De plus, nous ne souhaitons pas travailler par analyse des différentes parties constituant les CRH, car nous ne souhaitons pas développer une ontologie pour la rédaction et la seule analyse des CRH.

La liste des termes candidats et des sous-termes candidats est fournie par YATEA sous le format texte dans le fichier `nomFichierEntrée/default/raw/termList.txt` ou dans un fichier xml (extrait annexe J). La validation de cette liste a été réalisée manuellement par une personne non spécialiste du domaine. L'ETC a extrait 198 615 termes. Pour rendre l'analyse manuelle réalisable par une seule personne dans un temps donné, nous n'avons conservé que les termes de la liste avec une fréquence d'apparition supérieur à 4. Nous avons donc fait face à une liste de 27 744 termes en lien avec le domaine de la psychiatrie, à valider à la main.

Pour rendre cette validation manuelle réaliste dans le temps imposé par la thèse, nous avons procédé par élimination successive. Dans un premier temps nous avons supprimé automatiquement toutes les expressions numérales de la liste de termes à l'aide d'expressions régulières. Nous avons ensuite supprimé de la liste les noms de médicaments, les noms de troubles ou de symptômes, ou encore les noms relatifs à des parties du corps humains. Après ce premier travail, la liste à valider n'était plus que d'environ 5 000 termes.

4. La distance de Levenshtein calcul le nombre minimal d'opérations à effectuer pour transformer une chaîne 1 en une chaîne 2 et obtenir ainsi une mesure de similarité entre les deux chaînes.

Pour valider les termes en concepts, nous nous sommes appuyés sur l'organisation de la partie sociale de la SNOMED VF3.5. Elle contient 1 152 codes associés à un libellé et huit catégories principales : contexte sociale, états parentaux et civils, problèmes sociaux non causés par un trouble mentale, mode de vie, religions et philosophies, statuts économiques, établissement de soin et groupes ethniques et populations. Hormis la dernière catégorie, toutes étaient pertinentes pour appuyer la validation des termes. Nous avons donc regardé chacun des 5 000 termes restants dans la liste. Quand le terme correspondait à une catégorie sémantique de la SNOMED, il était validé. Nous obtenons à cette étape une liste de 1 100 concepts relatifs au domaine des facteurs sociaux et environnementaux des maladies psychiatriques.

### 5.4.2 Conceptualisation du domaine

Nous présentons ici le déroulement de la conceptualisation du domaine qui a mené aux résultats présentés. Tous les résultats de cette section ont été validés par des experts. La validation a permis une grande évolution de notre modèle de connaissances. Le chapitre 7 présente toutes ces évolutions plus en détail, dans la section 7.2.

Le première étape a été d'associer les concepts à leurs définitions et leurs équivalents lexicaux en langage naturel, qui correspond au nom du concept. Ensuite, nous avons commencé à développer la taxonomie, sous l'éditeur d'ontologie PROTÉGÉ [Musen, 2015]. Trois personnes spécialistes dans le développement d'ontologies, mais non spécialistes dans le domaine modélisé ont défini les quatre grandes classes de l'ontologie. Ces classes n'ont jamais changé malgré les ajustements effectués par la suite : (1) Attribut - adjectifs décrivant plus précisément les concepts ; (2) Concept de vie sociale - concepts sur les aspects sociaux (par exemple, l'éducation, la situation sociale ou la situation civile) ; (3) Être humain - pour représenter les êtres humains ; et (4) Groupes - pour formaliser le groupe primaire (la société, l'institution, le service ou la clinique) et secondaire (la famille). La création des relations de subsomption a été réalisée par la même personne qui avait validé les termes en concept. Puis en concertation avec les deux autres personnes qui avaient participé au choix des quatre grandes classes principales. Enfin, nous avons établi des relations entre les concepts, toujours en concertation entre les trois personnes qui ont développé les relations de subsomption. Cette taxonomie de relations visait à lier les concepts entre eux, grâce aux relations sémantiques autre que la relation de subsomption. Elle a beaucoup évolué depuis la première version de notre ontologie. Au début du développement, nous n'avions que 12 relations, maintenant nous avons 199 relations entre les concepts. Cet enrichissement est dû principalement à l'ajout de l'ontologie Family Health History Ontology (FHHO) présentée en section suivante.

Dans le tableau 5.2, nous remarquons que le nombre de sous-classes (2025) est supérieur au nombre de classes (1478). Cela s'explique en partie au fait que nous avons 172 concepts avec deux parents ou plus (seul les concepts de la classe « être humain » peuvent avoir plus d'un parent, par exemple une « mère adoptive » est une « mère », un « parent adoptif » et une « femme »). Cela ne correspond pas au principe différentiel. Deux raisons expliquent cette contradiction : (1) la Family Health History Ontology (FHHO) utilisée pour enrichir la section « être humain » contient plus de 150 concepts avec deux parents ; (2) certains concepts de l'ontologie devaient être représentés par deux concepts. Notre but est d'utiliser cette ontologie dans un système d'annotation. Il était donc plus facile

TABEAU 5.2 – Métriques du module « Facteurs sociaux et environnementaux des maladies psychiatriques »

Nombre de classes	1478
Nombre de relations	199
Profondeur maximale	11
Nombre d'enfants maximal	56
Nombre d'enfants moyens	3
Nombre de classes à enfant unique	146
Nombre de classes avec plus de 25 enfants	3
Nombre de classes non définie	1434
Nombre de sous classes	2025

pour nous de modéliser des concepts avec deux parents qu'avec une restriction logique, quand nous avons le choix. Par exemple, dans notre ontologie un « Établissement de santé » est un *lieu* et un *groupe primaire*.

TABEAU 5.3 – Quelques statistiques sur le module « Facteurs sociaux et environnementaux des maladies psychiatriques » (OntoPsychiaFSE). Les nombres de classes incluent les classes à deux parents.

Module social	Classes	Relations
OntoPsychiaFSE	1478 (172 avec deux parents)	199
Concept sur la vie sociale	1039	51
Être humain	560	144 (dont 140 issues de la FHHO)
Groupe	81	2

Dans le tableau 5.3, nous pouvons voir le nombre de concepts dans chacune des quatre classes principales. La classe « Concept sur la vie sociale » est le thème principal du domaine conceptualisé, de même que celui avec le plus de concepts. Nous remarquons alors que les concepts sous « être humain » sont très liés, ce qui s'explique par les liens entre les individus : les propriétés d'objet ne décrivent que des relations entre membres de la famille, environ 33 sont liées à des concepts de vie sociale et d'autres sont reliées à d'autres êtres humains, qui ne sont pas des membres de la famille, mais des amis. Nous ne présentons pas la classe réservée aux attributs, car ce n'est qu'une liste d'adjectifs.

Nous présentons ensuite les différentes classes enfants de la classe « Concept sur la vie sociale » :

1. **Changement - 35 concepts** : concepts relatifs à une transformation, une modification dans la vie d'une personne. Exemple : amélioration, déménagement, hospitalisation, changement d'activité professionnelle, reprise du travail, etc.
2. **Comportement - 48 concepts** : concepts qui décrivent une attitude, une manière d'être ou d'agir d'un individu. Exemple : attitude, désinvestissement, insertion sociale, inadaptation professionnelle, etc.
3. **Condition de vie - 141** : les concepts relatifs aux conditions de vie englobent (1) l'habitat (type : collectif, individuel, rural ou urbain ; qualité : insalubre ou non) ; (2) le logement (type : maison, foyer etc. ; qualité : salubre, précaire, etc.) ; (3) la situation



(type : financière, sociale ou judiciaire ; qualité : difficile, stressante, etc.) ; (4) l'environnement (isolé ou non isolé) ; (5) le mode de vie et (6) le statut dans la société.

4. **Caractéristique - 71 concepts** : ce qui marque la particularité de quelque chose ou de quelqu'un. Exemple : rang de naissance, fonctionnement au travail, capacité de socialisation, orientation sexuelle, etc.
5. **Droit et justice française - 195 concepts** : modélise l'organisation juridique et le personnel judiciaire en France. Classification des infractions et des peines selon le droit français.
6. **Éducation - 132 concepts** : modélise l'éducation scolaire et l'apprentissage intellectuel en France. Exemple : diplôme, instruction, niveau intellectuel, enseignement supérieur, manque d'apprentissage, redoublement, etc.
7. **Événement - 220 concepts** : concepts relatifs à un fait particulier qui se produit dans la vie d'une personne et qui s'inscrit dans une durée, qui a des conséquences. Certains de ces concepts pourraient être organisés sous le concept parent de « changement » (tel « séparation » ou « rencontre », mais l'importance est ici sur la notion de fait qui se produit, plus que sur la notion d'un changement). Exemple : accident, décès, naissance, échec, difficulté financière, séparation, rencontre.
8. **Lieu - 109 concepts** : ces concepts sont relatifs à un espace, un endroit physique. Exemple : maison de repos, hôpital, logement (qu'on retrouve également dans la branche des conditions de vie), laboratoire, etc.
9. **Relationnel - 133 concepts** : ces concepts expriment les liens que peut entretenir un individu dans la vie. Exemple : entente, mésentente, isolement, relation de dépendance, religion (est une relation d'ordre divin), etc.
10. **Sentiment - 26 concepts** : les concepts organisés sous la branche des sentiments, sont relatifs à des états affectifs, des émotions. Ces états affectifs sont divisés en états émotionnels négatifs, positifs ou variables. Exemple de concepts : angoisse, déception, solitude, bienveillance, sérénité, etc.

En annexe L, les figures L.2, L.1 et L.3 permettent de comprendre l'organisation conceptuelle de premier niveau de chacune des branches modélisées dans l'ontologie.

Une ontologie est également composée de relations, des propriétés d'objets, qui permettent de lier deux concepts entre eux et de formaliser la connaissance. Par exemple, la propriété « est ami de » relie deux individus entre eux. La figure 5.1 montre les deux premiers niveaux de la hiérarchie des propriétés d'objets contenues dans l'ontologie. Ces relations ne sont actuellement pas utilisées dans la définition des classes. Elles servent à décrire le modèle et les relations entre les concepts, tels que :

1. **a\_pour\_caractéristique** : cette relation relie un concept individu à un concept sous classe de la classe *Caractéristique*. Elle permet par exemple d'exprimer la relation « un individu a pour caractéristique un genre »
2. **a\_pour\_protecteur\_juridique** : permet de décrire la relation entre un individu et un curateur ou un tuteur.
3. **a\_pour\_relation** : permet de décrire les relations entre les individus, les liens de parenté aussi bien que les liens amicaux, entres autres.



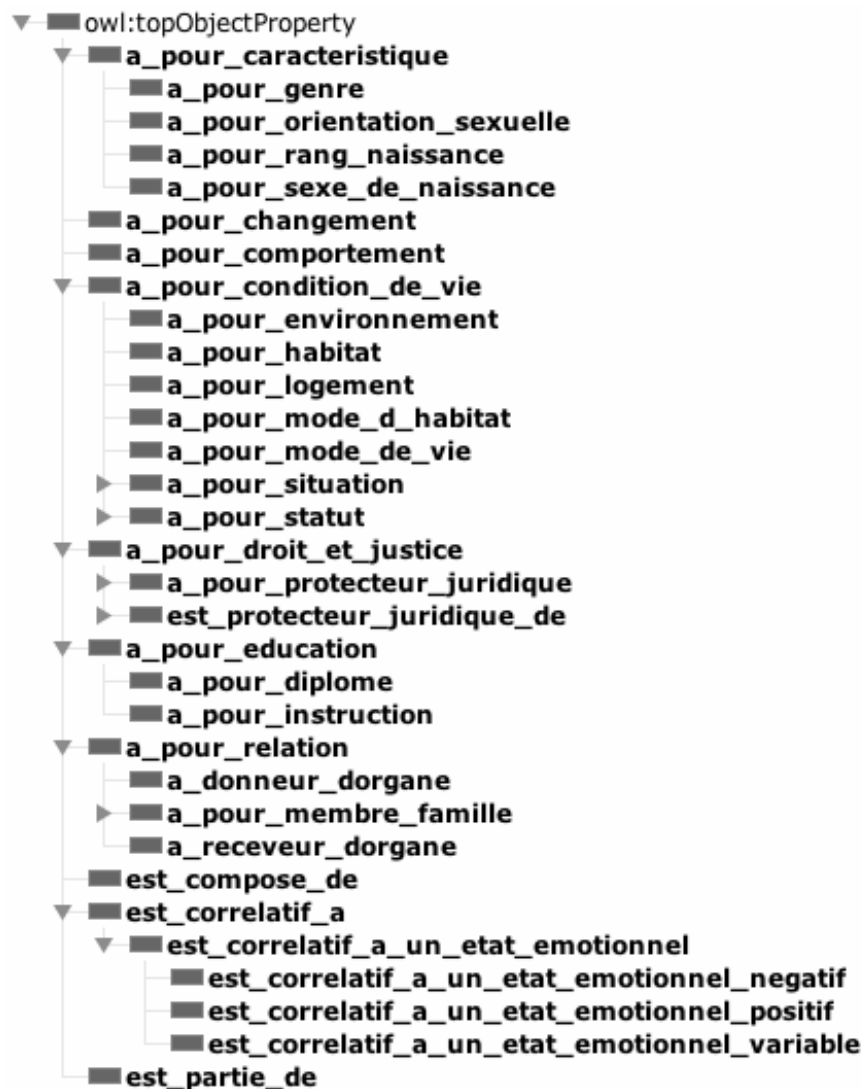


FIGURE 5.1 – Hiérarchie des relations modélisées dans l'ontologie pour décrire des liens entre les concepts.

4. **est\_correlatif\_a** : nous sert à exprimer la cause d'un état émotionnel négatif, positif ou variable. Actuellement, cette relation n'a pas de domaine défini. C'est-à-dire que nous n'avons pas fermé l'interprétation de la cause des états émotionnels.

Les relations de notre ontologie sont peu développées. Il est important à ce stade de description du modèle de rappeler qu'en l'état, notre modèle nous sert à décrire des facteurs sociaux ou environnementaux, qui peuvent ou non affecter la vie d'un patient, de manière positive, négative ou variable. Notre ontologie n'a donc pas de but computationnel. En d'autres termes, nous n'utilisons pas ce modèle pour faire des liens logiques entre les concepts ou pour déduire des informations, à partir des relations que les concepts entretiennent entre eux.

### 5.4.3 Enrichissement du domaine

Pour enrichir notre modèle, nous avons inclus la Family Health History Ontology (FHHO) de [Peace et Brennan \[2007\]](#), présentée en introduction du manuscrit. Elle conceptualise (1) les membres de la famille et les relations entre eux, (2) et les états de santé avec le diagnostic. Nous n'avons pas été intéressés par la deuxième partie de la FHHO, mais nous avons pris tous les concepts modélisant les membres de la famille, soit environ 140 concepts et presque autant de relations. Dans la FHHO, tous les membres sont définis sous un concept « structurant » (par exemple, « Adopted\_Cousin » (cousin adoptif) est une « Adopted\_Kin\_Relation » (relation adoptive). Nous avons voulu éviter ce système hiérarchique, nous avons donc renommé tous les concepts structurants, pour qu'ils soient des concepts relatifs à un individu (par exemple, « Adopted\_Kin\_Relation » (relation adoptive) est dans ONTOPSYCHIAFSE le concept « Adoptive\_Individual » (individu adopté). Nous n'avons pas modifié la hiérarchie conceptuelle de l'ontologie. Ces changements de nom de concept, n'ont pas servi les alignements qui ont été réalisés par la suite.

Nous avons réalisé un alignement avec l'outil ONAGUI (présenté en section 5.3.3), sur les libellés de la CIM-10 (section 1.3.2) qui contient un « Chapitre XXI : Facteurs influant sur l'état de santé et motifs de recours aux services de santé » composé de plus de 800 codes. Ainsi qu'un alignement sur la partie sociale de la SNOMED 3.5VF (section 1.3.1), sur laquelle nous nous étions appuyés pour la validation des termes, sans toutefois en faire un alignement, afin de rester indépendant de leur modélisation. L'outil a trouvé seulement 18 alignements potentiels de notre module avec la CIM-10 et 169 alignements potentiels avec la SNOMED 3.5VF. Nous avons alors validé 13 alignements avec la CIM-10 et 93 alignements avec la SNOMED 3.5VF. Les résultats obtenus sur la CIM-10 sont très mauvais. Ils peuvent s'expliquer par une utilisation très normée du langage dans la CIM-10. Ainsi que l'utilisation de phrases longues pour définir les concepts, ce qui ne permet pas de calculer correctement une mesure de similarité. Nous n'obtenons pas non plus de très bons résultats d'alignement avec la SNOMED 3.5VF. Ce qui peut s'expliquer par le choix des catégories intégrées à la SNOMED 3.5VF, par exemple tous ce qui concerne l'usage de drogue. Nous avons considéré que cette catégorie concerne les troubles et non la vie sociale des personnes. Nous avons également identifié certaines associations de termes qui ne répondaient pas à nos attentes : par exemple, « mendiant » et « fainéant » sont considérés comme synonymes dans la SNOMED 3.5VF. Par conséquent, nous avons décidé d'attendre une version française officielle de la SNOMED CT, avant d'affiner l'enrichissement (par alignement ou intégration) de concepts dans notre ontologie.

## 5.5 Synthèse

La modélisation des facteurs sociaux et environnementaux nous a confronté aux questions liées à la modélisation d'ontologies à partir de zéro. En effet, l'état de l'art a montré que les thésaurus et les ontologies médicales qui modélisent des aspects sociaux dans la vie d'un patient restent encore trop peu exploités en psychiatrie, donc aucun modèle existant n'a pu nous servir de base de développement. En outre, à l'heure actuelle, il n'existe aucun outil pour établir une possible corrélation entre la vie d'un patient et sa santé mentale.

Dans ce chapitre, nous avons présenté les différentes étapes de la conceptualisation

du module sur les facteurs sociaux et environnementaux. Nous avons suivi les étapes de la méthodologie hybride TOREUSE2ONTO [Drame, 2014] (présentée en section 3.4.2), du recueil et de l'anonymisation des données (présentée en annexe H) à la modélisation des connaissances, dans une ontologie. L'extraction des termes spécialisés de notre corpus a été réalisée en suivant la méthode TERMINAE (voir en section 3.2.2). Après validation des termes candidats, nous avons développé un module contenant 1478 concepts. Nous nous sommes inspirés de la méthode ARCHONTE [Bachimont *et al.*, 2002, Charlet *et al.*, 2006], et du principe différentiel, afin de discriminer les concepts entre eux, et de faire en sorte qu'un maximum de concepts soit modélisés avec un seul concept parent. Seul environ 250 concepts du module ne répondent pas à cet engagement. Enfin, la phase d'enrichissement du modèle a permis d'ajouter environ 140 concepts et presque autant de relations à partir de la FHHO, et d'aligner 93 concepts sur la SNOMED-3.5VF. Concernant les mauvais résultats d'alignement avec la CIM-10, il pourrait être intéressant d'intégrer les codes du « Chapitre XXI : Facteurs influant sur l'état de santé et motifs de recours aux services de santé » dans le futur, selon les indications des partenaires de l'hôpital Sainte-Anne.

La chapitre suivant présente la construction du module sur les maladies mentales.

# Chapitre 6

## Construction du module ontologique « maladies psychiatriques »

### Sommaire

<b>6.1 Choix du module</b>	<b>116</b>
6.1.1 Enjeux	116
6.1.2 Objectifs	116
<b>6.2 Méthode de construction des deux modules</b>	<b>117</b>
6.2.1 Étude du corpus : répartition des codes de la CIM-10 dans les CRH	117
6.2.2 Conceptualisation du modèle d'alignement des classifications	118
6.2.3 Conceptualisation du modèle de connaissances avec des termes extraits des CRH	118
<b>6.3 Résultats</b>	<b>121</b>
6.3.1 Étude du corpus : répartition des codes de la CIM-10 dans les CRH	121
6.3.2 Conceptualisation du modèle d'alignement des classifications	121
6.3.3 Conceptualisation du modèle de connaissances avec des termes extraits des CRH	128
<b>6.4 Les limites du module d'alignement des classifications</b>	<b>129</b>
<b>6.5 Synthèse</b>	<b>130</b>

*Nous venons de détailler le processus de développement du module d'ONTOPSYCHIA sur les facteurs de risque sociaux et environnementaux des maladies psychiatriques. Nous avons opté pour une approche hybride. La construction du modèle a été réalisée par approche ascendante, à partir de comptes rendus d'hospitalisation. Alors que l'enrichissement s'est fait par approche descendante, grâce à l'alignement et l'ajout de modèles déjà existants. Nous avons annoncé au chapitre précédent, que notre ontologie serait composée de deux modules. Nous présentons dans ce chapitre le module troubles/maladies psychiatriques. En première partie de la thèse, nous avons fait état d'un manque de modélisation formelle des classifications psychiatriques, et observé un manque de consensus autour des catégories descriptives des troubles psychiatriques. Nous nous sommes donc interrogés sur la manière de construire une ontologie qui tiendrait compte de ces problématiques. Quelles classifications modélisées et dans quels buts ? Doit-on tenir compte des règles de codage ou tendre vers une modélisation plus souple, qui laisserait place à la possibilité d'un consensus ? Pour la construction des modules sur les troubles psychiatriques, nous avons à nouveau opté pour une approche hybride. Pour la construction d'un module sur les classifications, nous partons d'une approche descendante. Nous avons réalisé un alignement des classifications CIM-10 et DSM-IV sur le modèle du DSM-5. Ensuite, nous avons construit un modèle par méthode ascendante (de la même manière qu'au Chapitre précédent), à partir de l'extraction de termes candidats extraits des CRH de l'hôpital Sainte-Anne. La première section de ce chapitre présente le module, la deuxième section la méthode de développement utilisée, la troisième section présente les résultats et une quatrième section discute les limites du module.*

## **6.1 Choix du module**

### **6.1.1 Enjeux**

En introduction de ce manuscrit, nous avons indiqué qu'environ une personne sur trois souffrira d'un trouble mental au cours de sa vie. Malgré des classifications des troubles mentaux internationalement reconnues (voir la section 1.3), la catégorisation des patients selon des critères diagnostiques reste problématique. En effet, les troubles sont dénotés par des syndromes qui possèdent des symptômes propres à une ou plusieurs catégories diagnostiques. Lorsqu'un patient présente des symptômes qui franchissent les frontières des catégories descriptives et qui ne correspondent à aucune des sous-catégories diagnostiques, le recours au « NOS » *Not Otherwise Specified* permet d'attribuer une étiquette diagnostique et de ne pas laisser le patient exempt des codages. Cependant, ce type de recours marque une incohérence entre la description des troubles répertoriés dans les classifications et la réalité clinique, telle qu'elle est appréhendée par les professionnels de santé mentale.

### **6.1.2 Objectifs**

Dans l'introduction et la sections 1.3, nous avons fait état d'une faiblesse dans la modélisation des classifications psychiatriques, par les acteurs de l'ingénierie des connaissances. Les rares ontologies qui modélisent des classifications médicales, telles que le DSM ou la CIM ne sont généralement pas disponibles librement. En outre, les

professionnels en santé mentale disposent de différents systèmes de classification, pour les aider dans leur diagnostic et il est rare que leur utilisation se restreigne à un seul de ces systèmes. Pourtant, rien ne leur permet formellement de faire un lien entre deux codages diagnostics, alors que les systèmes de codage économique obligent, du moins en France, à l'utilisation d'un système bien précis, actuellement la CIM-10. Nous souhaitons proposer un modèle pour favoriser l'interopérabilité des données contenues dans les CRH. Ce modèle à destination du Service-Hospitalo Universitaire de Sainte-Anne permettra d'avoir accès aux catégories diagnostiques de trois grandes classifications psychiatriques. Le but est de disposer d'une ressource OWL, qui permette l'interopérabilité entre les diagnostics et l'indexation sous différents systèmes de codage.

Ensuite, nous souhaitons développer un modèle de connaissances avec des termes extraits des CRH, qui sont le reflet de la réalité clinique. En effet, notre corpus contient un grand nombre d'informations médicales. Ces informations peuvent éclairer sur l'utilisation du vocabulaire psychiatrique et sur les concepts manipulés dans ce domaine. L'objectif à long terme est de proposer une alternative aux systèmes de classification actuels et de réduire l'incohérence entre la description des troubles et la réalité clinique. Cependant, nous avons vu lors de la présentation du projet Research Domain Criteria (RDoC) en section 1.3.5, que le développement de nouvelles méthodes pour classer les troubles mentaux est un chantier immense, qui demande l'implication d'un grand nombre de chercheurs. Nous ne disposons pas, dans le cadre de notre projet, des ressources humaines pouvant aboutir à ce genre de projet de grande envergure. Par ailleurs, bien que l'ambition affichée de RDoC soit de proposer un nouveau système de classification des troubles psychiatriques, la matrice n'est pas une classification à l'heure actuelle. Elle est à l'état de description des résultats issus de la recherche en cognition, biologie, psychiatrie, etc. En outre, les auteurs de la matrice RDoC n'ont pour le moment décrit aucun plan, aucune recommandation, pour utiliser RDoC comme une classification, ou au moins nourrir ses résultats pour l'utilisation des classifications actuelles. Cette absence de perspective révèle le niveau de difficulté à établir un nouveau modèle de classification en psychiatrie, qui fera consensus au sein de la communauté. Elle met également en avant que les résultats bruts issus de la matrice ne sont pas exploitables, pour la construction d'un module ontologique. Nous avons donc concentré nos efforts sur le développement par approche descendante, d'un modèle fondé sur les classifications actuelles.

## 6.2 Méthode de construction des deux modules

### 6.2.1 Étude du corpus : répartition des codes de la CIM-10 dans les CRH

Avant de réaliser un alignement des classifications, nous avons souhaité étudier l'étendue de l'utilisation de la codification CIM-10, au sein des CRH de l'hôpital Sainte-Anne. Cette démarche a pour but de mettre en évidence les besoins d'alignement et d'enrichissement des catégories diagnostiques les plus utilisées. Pour cette étude, nous nous sommes appuyés sur la version ontologique de la CIM-10 réalisée par l'équipe du projet qui a pour objectif le Catalogage et l'Indexation des Sites Médicaux de langue Française

(CISMeF) accessible sur le site du projet<sup>1</sup> ainsi que sur BioPortal<sup>2</sup>.

### 6.2.2 Conceptualisation du modèle d'alignement des classifications

Selon les recommandations et avis de l'équipe de l'hôpital Sainte-Anne, nous avons choisi comme modèle de base pour l'alignement, la dernière version de l'édition 5 du DSM (présentée en section 1.3.3). Nous avons aligné sur cette classification le DSM IV TR (qui n'est autre que le texte révisé du DSM IV) et la CIM-10, qui sont les deux autres classifications en usage à l'hôpital. Le DSM 5 s'inscrit dans une approche dimensionnelle (présentée en section 1.3). Ce type d'approche tend à décrire les troubles au travers de spectres, tels que le « spectre de troubles schizophréniques » ; le « spectre des troubles bipolaires » ou encore le « spectre autistique ». Ce concept dénote la diversité des troubles et des symptômes qui peuvent se retrouver dans une pathologie et l'axe sur lequel les patients évoluent. L'approche dimensionnelle inclut également la notion de seuil pathologique ou de mesure des troubles, qui va du normal au pathologique. Le but de ce genre d'approche est de réduire les comorbidités, les chevauchements entre plusieurs catégories de troubles, le recours au « NOS » ou encore la surmédication ou prescription cumulative (le fait de prescrire une molécule pour chaque symptôme ou trouble). L'approche dimensionnelle suscite également des inquiétudes, nous les avons abordées brièvement en section 1.3. En effet, si nous pensons les troubles en termes de degrés et non plus en termes de syndromes, le seuil pathologique devient la mesure du trouble. Allan Frances parle ainsi de « médicalisation de la normalité », face à l'inquiétude de prescription pour des symptômes bénins, qui ne justifient pas de traitements médicamenteux. Le risque, si les seuils sont mal évalués, est de voir un nombre important de faux positifs diagnostiqués avec un trouble psychiatrique.

L'alignement des classifications suit donc la catégorisation du DSM 5. Pour réaliser cet alignement, nous nous sommes référés aux indications fournies par les classifications elles-mêmes. En effet, certains codes de la CIM-10 sont alignés sur les codes du DSM 5. Nous avons également utilisé le logiciel ONAGUI présenté en section 5.3.3, que nous avons déjà utilisé lors de la construction du module sur les facteurs sociaux et environnementaux.

### 6.2.3 Conceptualisation du modèle de connaissances avec des termes extraits des CRH

L'enrichissement du modèle est réalisé par approche ascendante. Il reprend la méthodologie présentée au Chapitre 5, qui nous a permis de construire le module sur les facteurs sociaux et environnementaux.

Toutefois, en 2015, soit deux ans après le début de notre étude, un nouvel extracteur de termes candidats, BIOTEX, a été développé. Nous l'avons testé sur notre corpus et il a offert de meilleures performances que YATEA (voir en section 3.2.4). Nous avons choisi de poursuivre le développement de notre ontologie sur les résultats d'extraction de BIOTEX. En effet, l'analyse des données brutes est notre matériel de construction premier. Les experts interviennent par la suite, lors de la validation du modèle, quand ce dernier est déjà

1. <http://www.chu-rouen.fr/cismef/projet-cismef/a-propos/>

2. <http://umls.biportal.lirmm.fr/ontologies/CIM-10?p=classes>

construit. Un bon taux de rappel dans la tâche d'extraction des termes est essentiel, pour proposer aux experts une couverture conceptuelle maximale du domaine.

### Extraction de termes candidats avec BIOTEX

Pour réaliser la suite de notre ontologie, nous nous sommes intéressés au logiciel BIOTEX. Pour résumer ce qui a été détaillé en section 3.2.3, BIOTEX utilise quatre méthodes pour déterminer l'importance d'un terme : (1) l'annotation morpho-syntaxique du corpus ; (2) l'extraction des termes candidats à l'aide d'une liste de motifs/patrons linguistiques ; (3) le classement des termes selon leur pertinence, appuyé par diverses méthodes statistiques ; (4) le calcul des cooccurrences des termes candidats, afin d'améliorer le classement final. Le logiciel BIOTEX se présente comme une application Web qui peut-être utilisée directement en ligne à cette adresse : <http://tubo.lirmm.fr/biotex/index.jsp> ou bien en librairie JAR (Java) utilisable sur machine. Dans le cas de l'application Web, le fichier de sortie fournit une liste maximum de 1200 termes candidats classés par ordre de pertinence. Dans le cas de la librairie Java, BIOTEX fournit quatre fichiers de sortie : liste des *unigram*, *bigram*, *3 gram*, *4 gram* et *plus*, ainsi qu'un fichier *ALL\_gram* contenant tous les termes extraits. L'utilisateur peut jouer sur le nombre minimal de fréquence des termes à analyser, afin d'optimiser le traitement de son corpus. Dans le cadre de l'analyse de gros corpus cela permet d'améliorer le temps de traitement. BIOTEX intègre l'annotation morpho-syntaxique via l'annotateur TREETAGGER. La seule chose que l'utilisateur ait à faire est donc de charger son corpus en .txt dans le logiciel, sélectionner les paramètres (par exemple, concernant la langue, la fréquence minimale, ou la mesure statistique) et patienter (parfois quelques heures).

Pour illustrer un exemple d'analyse avec BIOTEX, reprenons le même exemple « Trouble de l'adaptation avec humeur anxiodépressive chez un patient de 40 ans, souffrant d'un trouble de la personnalité. » :

En figure 6.1, le logiciel a identifié sept termes candidats. Il a réalisé pour nous une pré-validation en nous proposant de valider deux termes qu'il a pu retrouver dans le MESH<sup>3</sup>. Les autres termes candidats sont annotés en « Don't know » et nous n'avons plus qu'à faire le choix de les valider ou les invalider. Le logiciel fournit également un fichier XML téléchargeable, contenant les résultats de l'extraction. Un extrait de ce fichier est visible en annexe K.

---

3. Rappel : Medical Subject Headings fait office de thésaurus de référence dans le domaine biomédical.



The screenshot displays the BioTex web application interface. At the top, the logo 'BioTex' is followed by the tagline 'BIOmedical Term EXtraction'. A navigation bar includes links for 'Extraction', 'Evaluation' (which is highlighted), 'About and Citation', 'JAR and Documentation', and 'Contact Us'.

The main section is titled 'List of Extracted Terms'. It features a dropdown menu set to 'All' and a 'Download XML File' button. Below this, it indicates 'Number of terms : 7' and 'Page 1 of 1'.

A table lists the extracted terms with columns for 'N°', 'Extracted Term', and 'Biomedical term? Y/N/DK'. The terms are evaluated as follows:

N°	Extracted Term	Biomedical term? Y/N/DK
1	trouble de l'adaptation avec humeur anxiodépressive	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
2	trouble de l'adaptation avec humeur	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
3	adaptation avec humeur anxiodépressive	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
4	trouble de l'adaptation <i>Validate by : MeSH</i>	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
5	trouble de la personnalité <i>Validate by : MeSH</i>	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
6	adaptation avec humeur	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW
7	humeur anxiodépressive	<input checked="" type="checkbox"/> YES <input type="checkbox"/> NO <input type="checkbox"/> DON'T KNOW

At the bottom of the table, it again shows 'Number of terms : 7' and 'Page 1 of 1'.

On the right side, there are sections for 'Institutions' (listing 'Laboratoire Informatique Robotique Microelectronique Montpellier' with logos for 'UNIVERSITÉ MONTPELLIER', 'cnrs', and 'TETIS') and 'Sponsors' (listing 'SIFR project').

At the very bottom, the copyright notice reads '© Copyright Juan Antonio Lossio Ventura.' and a link to 'SIFR Project' is provided.

FIGURE 6.1 – Exemple d’une extraction de termes candidats avec le logiciel BIOTEX.

Cette extraction a été réalisée à partir de la phrase « Trouble de l’adaptation avec humeur anxiodépressive chez un patient de 40 ans, souffrant d’un trouble de la personnalité. »

## 6.3 Résultats

### 6.3.1 Étude du corpus : répartition des codes de la CIM-10 dans les CRH

Pour réaliser cette étude, nous avons extrait automatiquement les codes issues de la classification CIM-10 des CRH, à l'aide d'expressions régulières. Au préalable, nous avons réalisé une correction semi-automatique des codes mal formés ou mal orthographiés du type F2000 ou F.200 pour F20.00. Ensuite, à chaque code, nous avons associé son libellé dans la CIM-10 version française si celui-ci existait. S'il était absent de la CIM-10 version française, nous allions voir dans la CIM-10 version anglaise<sup>4</sup> si le code existait. Nous comptons sept codes qui se référaient à la CIM-10 anglaise. Ces codes apparaissent entre une à deux fois dans l'ensemble des CRH. Cependant, ces codes sont certainement le résultat d'erreurs de saisies. En effet, les CRH ont été rédigés à une date antérieure à la version 2017 de la CIM-10 et nous ne sommes pas parvenus à trouver ces codes dans les versions de la CIM-10 correspondant à la période de rédaction des CRH. Enfin, dans les cas où nous ne pouvions trouver le code équivalent ni dans la CIM-10 française, ni dans sa version anglaise, nous avons laissé la valeur nulle pour conserver le code néanmoins. Les 43 codes concernés par cette manipulation n'apparaissaient jamais plus de trois fois dans l'ensemble des CRH.

TABLEAU 6.1 – Répartition des codes selon leur fréquence d'apparition dans les CRH.

Nombre de codes	Intervalle de fréquence
6	[200 ; 686]
30	[50 ; 200]
62	[10 ; 50]
258	< 10

Pour rappel, la classification CIM-10 contient plus de 1300 codes propres au domaine de la psychiatrie. Dans l'ensemble des CRH nous avons extrait 356 codes CIM-10 différents dans 7 756 CRH sur environ 8 700. De facto, un peu plus de 1 000 CRH n'étaient pas codés avec la CIM-10. Nous comptons 258 codes qui apparaissent moins de dix fois dans l'ensemble des CRH, 62 codes qui apparaissent entre dix et 50 fois, 30 codes qui apparaissent entre 50 et 200, et six codes qui apparaissent entre 200 et 686 fois. Nous avons répertorié les 20 codes les plus fréquents dans le tableau 6.7.

### 6.3.2 Conceptualisation du modèle d'alignement des classifications

Après discussion avec les responsables du projet à l'hôpital Sainte-Anne, nous avons sélectionné trois classifications de référence à aligner. Ces classifications sont celles en usage à l'hôpital Sainte-Anne : le DSM 5, le DSM IV TR et la CIM-10 (extrait en annexe C). Nous avons utilisé la version OWL de la CIM-10 réalisée par le CISMEEF, la version OWL du Mini DSM 5 version française réalisée par Xavier Aimé, et la version OWL du DSM IV TR réalisée pour ce travail. Nous avons ensuite procédé à l'alignement des classifications,

4. La version anglaise de la CIM-10 2017 est consultable en ligne à cette adresse : [ftp://ftp.cdc.gov/pub/Health\\_Statistics/NCHS/Publications/ICD10CM/2017/](ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2017/)

en trois étapes. En outre, nous observons que ces classifications se présentent différemment : la CIM-10 contient des codes chiffrés associés à un libellé en langage naturel (voir annexe C), alors que les DSM ne contiennent pas de code, uniquement une liste des catégories associées à des informations diagnostiques. Pour la suite, nous parlerons de codes pour la CIM-10 et de catégories pour les DSM.

### Alignement selon les indications du DSM 5

La cinquième version du DSM, tout comme la quatrième version, contient un alignement avec les codes de la CIM-10. Cet alignement est réalisé sur environ 500 catégories diagnostiques et 58 codes de la CIM-10 font partie intégrante de la hiérarchie conceptuelle du DSM 5. Une annotation « CIM\_10Id » associée au code CIM-10 a été ajoutée à toutes les catégories du DSM 5 alignées selon les recommandations indiquées par le manuel. Le tableau 6.2 présente les résultats.

TABEAU 6.2 – Résultats de l'alignement selon le DSM 5.

	CIM-10
DSM 5	565

### Alignement du DSM IV TR et enrichissement de l'alignement CIM-10 avec l'outil ONAGUI

Lors de la deuxième étape, nous avons procédé à l'alignement du DSM IV TR sur les catégories du DSM 5 à l'aide de l'outil ONAGUI. Pour rappel, cet outil utilise les techniques de mesures de similarité entre les chaînes de caractères, afin d'opérer un alignement lexical entre deux termes ou suites de termes. Nous en avons profité pour réaliser le même type d'alignement avec la CIM-10, afin d'enrichir l'alignement proposé par le DSM. Cette étape nous a permis d'aligner 144 catégories du DSM IV TR (via l'annotation « DSM\_4Id ») et d'ajouter 16 codes de la CIM-10 à l'alignement.

Les ajouts de codes issus de la CIM-10 sont pour moitié le résultat de corrections d'annotations. En effet, certains codes étaient mal annotés dans l'ontologie. Ils apparaissaient par exemple sous le label *alternative label* ou avec l'annotation réservée aux codes issus de la CIM 9 (une autre version de la CIM, mais que nous n'utilisons pas dans nos travaux). Le compte des annotations CIM-10, pour dénombrer les codes alignés sur les catégories du DSM 5, était donc erroné. Ces codes mal annotés n'apparaissaient pas dans les résultats en tant que code CIM-10, malgré leur présence dans l'ontologie. Les autres huit ajouts concernent essentiellement des alignements de titre de catégories descriptives, qui ne sont pas répertoriés dans le manuel. Par exemple, la catégorie « Handicap Intellectuel » correspond aux codes CIM-10 « F70-F79 », la catégorie « Tic » correspond au code F95, ou encore la catégorie « Troubles des apprentissages » correspond au code F81. Le tableau 6.3 présente les résultats.

TABEAU 6.3 – Résultats de l'alignement avec OnaGUI.

	DSM IV TR	CIM-10
DSM 5	144	581 (+16)

### 6.3 Résultats

TABLEAU 6.5 – Répartitions des alignements dans le modèle de connaissances des classifications.

Classes issues du DSM 5 (nombre de concepts)	Nombre de concepts alignés sur le DSM IV TR	Nombre de concepts alignés sur la CIM-10	Nombre de concepts alignés sur les deux classifications
Autre situation pouvant faire l'objet d'un examen clinique (195)	22	141	15
Trouble des mouvements et autres effets indésirables induit par un médicament (22)	6	16	6
Trouble mental (618)	306	431	269

#### Alignement du DSM IV TR par transitivité avec l'alignement de la CIM-10

La quatrième édition du DSM contient 481 catégories alignées sur la CIM-10. Nous avons décidé d'exploiter cet alignement pour enrichir l'alignement du DSM IV TR sur le DSM 5. Nous avons considéré l'hypothèse transitive suivante :

Si (Code DSM IV = Code CIM-10) et (Code CIM-10 = Code DSM 5)  
alors (Code DSM IV = Code DSM 5)

Cette opération a permis un réel enrichissement de l'alignement. Nous avons ajouté 188 catégories du DSM IV TR alignées sur les catégories du DSM 5. Nous avons également constaté que 7 codes de la CIM-10 n'étaient pas encore alignés, selon les recommandations. Le tableau 6.4 présente les résultats.

TABLEAU 6.4 – Résultats de l'alignement transitif.

	DSM IV TR	CIM-10
DSM 5	332 (+188)	588 (+7)

#### Codes alignés sur les trois classifications

Nous présentons l'alignement des trois classifications dans le diagramme 6.2. Nous avons aligné 290 codes sur les trois classifications. Ce qui peut sembler peu, étant donné l'étendue des classifications. Néanmoins, ce nombre est important si nous tenons compte des grandes disparités de conceptualisation des troubles entre les classifications.

La hiérarchie conceptuelle du modèle est organisée selon l'arborescence du DSM 5, nous retrouvons donc les grandes catégories du DSM 5. Nous illustrons dans le tableau 6.5, les trois classes principales du modèle avec le nombre d'alignements sur le DSM IV TR et le DSM 5 qu'elles contiennent. La figure 6.3 illustre la façon dont les catégories issues d'une classification deviennent des classes d'une ontologie, organisées par la relation de subsumption. Elle permet également de visualiser l'alignement via les annotations.

Cet alignement doit encore être validé par des professionnels de santé mentale. En effet, certains alignements sont difficiles à valider, en particulier pour un non professionnel du domaine. Par exemple, le code F10.20 de la CIM-10 qui correspond au libellé « Troubles mentaux et du comportement liés à l'utilisation d'alcool, syndrome de dépendance actuellement abstinent » correspond selon le manuel du DSM IV TR à la catégorie

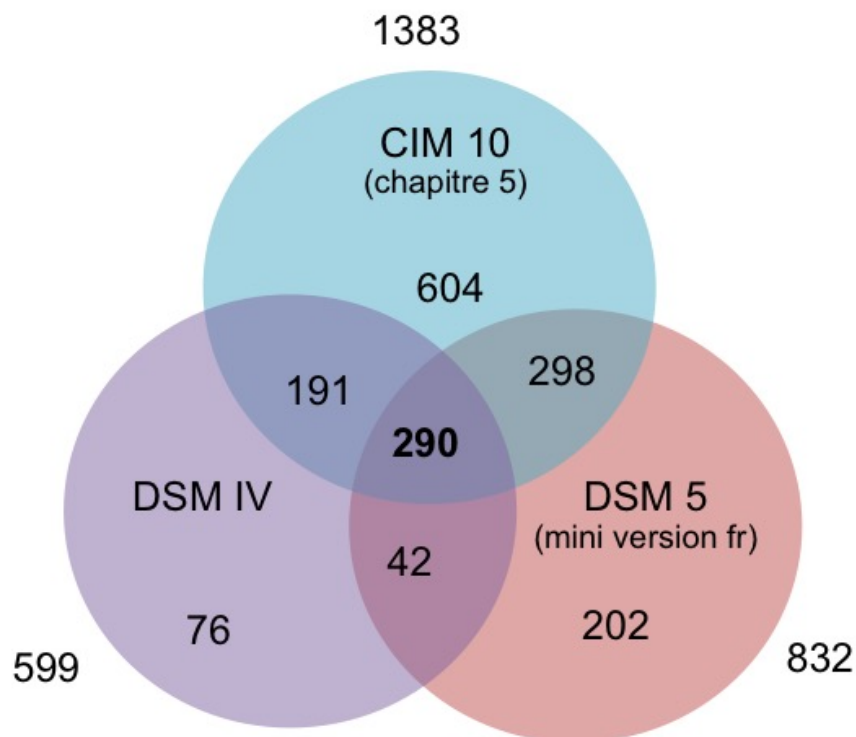


FIGURE 6.2 – Diagramme des alignements entre CIM-10, DSM IV TR et DSM 5.

Les chiffres correspondent aux nombres de concepts présents dans les classifications et aux résultats des alignements.

Class hierarchy: Schizophrenie

owl:Thing

- AutreSituationPouvantFaireLObjetDUnExa
- TroubleDesMouvementsEtAutresEffetsInd
- TroubleMental
  - AutreTroubleMental
  - DysfonctionSexuelle
  - DysphorieDeGenre
  - SpectreDeLaSchizophrenieEtAutreTrou
    - AutreTroubleDuSpectreDeLaSchizop
    - CatatonieAssocieeAUnautreTrouble
    - CatatonieNonSpecifiee
    - Schizophrenie
    - TroubleCatatoniqueDuAUneAutreAff
    - TroubleDelirant
    - TroubleDuSpectreDeLaSchizophreni
    - TroublePsychotiqueBref
    - TroublePsychotiqueDuAUneAutreAff
    - TroublePsychotiqueInduitParUneSub
    - TroubleSchizoaffectif
    - TroubleSchizophreniforme
  - TroubleAnxieux
  - TroubleASymptomatologieSomatiqueEt
  - TroubleBipolaireEtApprente
  - TroubleDeLaConduiteAlimentaireEtDeL
  - TroubleDeLALeteranceVeilleSommeil
  - TroubleDeLaPersonnalite
  - TroubleDepressif
  - TroubleDisruptifDuControleDesImpulsio
  - TroubleDissociatif
  - TroubleDuControleSphincterien
  - TroubleLieADesTraumatismesOuADEsF
  - TroubleLieAUneSubstanceEtTroubleAd
  - TroubleNeurocognitif
  - TroubleNeuroDeveloppemental
  - TroubleObsessionelCompulsifEtAppren
  - TroubleParaphilique

Asserted

Annotations: Schizophrenie

Annotations

- skos:prefLabel

[language: fr]

schizophrénie
- CIM\_10Id

[type: xsd:string]

F20.9
- CIM\_9Id

[type: xsd:string]

295.90
- DSM\_4Id

[language: fr]

Schizophrénie type indifférencié

Description: Schizophrenie

Equivalent To

+

SubClass Of

+

SpectreDeLaSchizophrenieEtAutreTroublePsychotique

General class axioms

+

FIGURE 6.3 – Hiérarchie conceptuelle issue de la catégorisation du DSM 5 dans la fenêtre de gauche. Exemple d'alignement de code avec la CIM 10 et de catégorie avec le DSM IV TR à droite.

«Dépendance alcoolique en rémission prolongée complète/partielle » et correspond dans le DSM 5 à un « Trouble léger/moyen/sévère de l'usage de l'alcool ». Nous constatons ainsi des différences dans les labellisations des troubles, qui nécessitent une expertise médicale des symptômes associés.

Pour illustrer les différences entre les classifications, nous pouvons prendre les codes et catégories qui mentionnent la « schizophrénie » dans les trois classifications. Le DSM IV TR contient 16 catégories, le DSM 5 en contient 17 et la CIM-10 contient 126 codes. Néanmoins, seul 10 catégories sont alignées sur les trois classifications.

TABLEAU 6.6 – Alignement des codes et catégories relatifs à la « schizophrénie ».

Catégorie / Libellé DSM 5	DSM IV TR	CIM-10
Schizophrénie	Schizophrénie type indifférencié	F20.9
Trouble catatonique du à ...	trouble catatonique du à ...	F06.1
Trouble délirant	trouble délirant	F22
Trouble du spectre de la schizophrénie NS	Trouble psychotique NS	F29
Trouble psychotique bref	Trouble psychotique bref	F23
Trouble psychotique du à une autre affection médicale avec hallucinations	Trouble psychotique dû à ... [Indiquer l'affection médicale générale] avec hallucinations	F06.0
Trouble psychotique du à une autre affection médicale avec idées délirantes	Trouble psychotique dû à ... [Indiquer l'affection médicale générale] avec idées délirantes	F06.2
Trouble schizoaffectif de type bipolaire	Trouble schizo-affectif type bipolaire	F25.0
Trouble schizoaffectif de type dépressif	Trouble schizo-affectif, type dépressif	F25.1
Trouble schizophréniforme	Trouble schizophréniforme	F20.81

### Mise en parallèle des résultats de l'alignement avec notre corpus

Nous avons extrait la liste des 20 codes CIM-10 les plus fréquemment rencontrés dans les CRH et nous les avons confrontés à notre modèle. Le tableau 6.7 présente les résultats.

Pour les 20 codes CIM-10 les plus fréquents, nous comptons six codes qui ne sont pas alignés sur les trois classifications. Ces codes correspondent principalement à la schizophrénie, puisque cinq codes sur six sont concernés par ce défaut d'alignement. Nous avons pourtant calculé précédemment que dix catégories sur 16 que contient le DSM 5 concernant la schizophrénie étaient alignées sur les deux autres classifications. Cette lacune met en avant les difficultés d'alignement entre deux systèmes de classification différents.

TABLEAU 6.7 – Liste des 20 codes CIM-10 les plus fréquemment rencontrés dans les CRH de Sainte-Anne, alignés sur les catégories des DSM.

Nb	Code CIM-10	Code DSM 5	Code DSM IV TR
686	F20.0		Schizophrénie, forme paranoïde, à évolution continue
643	F20.3		Schizophrénie, type paranoïde, épisodique sans symptômes résiduels entre les épisodes

Continue page suivante...

### 6.3 Résultats

TABLEAU 6.7 ...suite de la page précédente

Nb	Code CIM-10	Code DSM 5	Code DSM IV TR
362	F32.2	Trouble dépressif caractérisé avec épisode isolé grave	Trouble dépressif majeur, épisode isolé, sévère sans caractéristiques psychotiques
296	F31.4	Épisode dépressif grave	Trouble bipolaire I, épisode le plus récent dépressif, sévère sans caractéristiques psychotiques
229	F33.2	Trouble dépressif caractérisé avec épisode récurrent grave	Trouble dépressif majeur, récurrent, sévère sans caractéristiques psychotiques
215	F25.0	Trouble schizoaffectif de type bipolaire	Trouble schizo-affectif type bipolaire
194	F10	Trouble lié à l'alcool	Troubles liés à l'alcool
178	F31.2	Trouble bipolaire I, épisode le plus récent maniaque, sévère avec caractéristiques psychotiques	épisode maniaque avec caractéristiques psychotiques
159	F25.1	Trouble schizoaffectif de type dépressif	Trouble schizo-affectif, type dépressif
160	F20.1 (Schizophrénie hébéphrénique)		
129	F33.1	Trouble dépressif caractérisé avec épisode récurrent moyen	Trouble dépressif majeur, récurrent, moyen
125	F25.9 (Trouble schizo-affectif, sans précision)		
115	F32.3	Trouble dépressif caractérisé avec épisode isolé et caractéristiques psychotiques	Trouble dépressif majeur, épisode isolé, sévère avec caractéristiques psychotiques
108	F25.2 (Trouble schizo-affectif, type mixte)		
103	F31.1		Trouble bipolaire I, épisode le plus récent maniaque (léger/moyen/sévère)
100	F60.31	Personnalité émotionnellement labile type borderline	Personnalité borderline
98	F20 (Schizophrénie)		
95	F60.9	Trouble de la personnalité non spécifié	Trouble de la personnalité NS
87	F31.3		Trouble bipolaire I, épisode le plus récent dépressif, léger/moyen
135	F32.1	Trouble dépressif caractérisé avec épisode isolé moyen	Trouble dépressif majeur, épisode isolé, moyen



### 6.3.3 Conceptualisation du modèle de connaissances avec des termes extraits des CRH

Nous reprenons ici la même méthode que celle mise en œuvre au chapitre précédent, pour la construction du module sur les facteurs sociaux et environnementaux (voir les sections 5.3.1 et 5.3.2).

#### Extraction de termes candidats avec BIOTEX

Les résultats de BIOTEX sont classés par ordre de pertinence, afin de limiter l'analyse manuelle. Nous avons retenu de notre expérience avec YATEA, qu'une liste de plus de 20 000 termes pour un validateur non spécialisé était certes, fastidieuse à valider, mais faisable en temps humain de deux à quatre mois complets. Le validateur n'a jamais travaillé à temps complet sur la validation des termes, nous ne pouvons donc fournir de temps de validation précis. De plus, ce temps peu varier selon le domaine d'expertise et les connaissances propres du validateur sur ce domaine.

BIOTEX a extrait environ 200 000 termes candidats, tout grammes confondus. Nous avons sélectionné les 5 000 premiers termes candidats de chacun des quatre fichiers résultats (bigram, 3gram, 4gram, et plus), pour réduire à 20 000 termes la liste à valider manuellement.

#### Validation des termes en concepts

Ce travail de validation des termes candidats en concepts s'est échelonné sur plusieurs mois, en parallèle à la validation du module sur les facteurs sociaux et environnementaux. Une seule personne non spécialiste du domaine a travaillé seule sur la validation d'un peu plus de 5 500 termes médicaux. Lors d'une seconde validation, la même personne a isolé 2 711 termes propres à la psychiatrie. Ensuite, des scripts écrits en langage Perl ont permis le formatage automatique des termes en classe OWL d'une ontologie. Enfin, les concepts ont été organisés entre eux, pour avoir à la fin de ce travail, une taxonomie de 2 499 concepts. Les images M.1, M.2 et M.3 sont des captures d'écran du modèle de connaissances présenté sous le logiciel Protégé. Elles permettent de comprendre l'organisation conceptuelle de premier niveau des branches modélisées.

Nous n'avons pas eu l'occasion de discuter de notre modélisation avec les acteurs du domaine de la psychiatrie. Les principales catégories conceptuelles de notre organisation (parcours de soin, acte médical, diagnostic, etc.) sont donc pour le moment relativement intuitives. L'organisation des concepts enfants de ces catégories est basée sur un rapprochement lexical des termes (un « trouble de la personnalité complexe » est un « trouble de la personnalité » qui est lui-même un « trouble ») qui est le résultat de l'extraction des termes. Pour éclairer notre choix de conceptualisation, nous décrivons chaque branche :

1. **Acte médical - 23 concepts** : cette branche regroupe les concepts relatifs à des consultations ou des examens médicaux.
2. **Diagnostic - 25 concepts** : diagnostic de bipolarité, de démence, de dépression, et toutes autres expressions lexicales de type *diagnostic + trouble*.
3. **État - 237** : regroupe des concepts relatifs à un état mental, émotif, psychiatrique, somatique, etc. C'est-à-dire toutes expressions de type *Etat + qualificatif relatif à une observation médicale*.

4. **Observation médicale - 1750 concepts** : cette branche contient le plus grand nombre de concepts. Nous avons donc créé trois sous classes pour décrire les concepts plus finement :
  - (a) **Observation médicale sur la condition du patient - 334 concepts** : nous avons regroupé les concepts qui semblent définir la condition du patient tels que l'errance, la désorganisation, la mélancolie, la présentation, le tendance, le trait de personnalité, le vécu, etc.
  - (b) **Observation médicale sur l'évolution de l'état du patient - 401 concepts** : ces concepts sont relatifs à une évolution dans les symptômes, troubles, traitements, ou toutes autres faits médicaux, selon que cette évolution soit croissante (augmentation de l'appétit, augmentation de la posologie, développement des idées délirantes, etc.), décroissante ( baisse de plaisir, diminution de l'anxiété, régression des troubles, etc.), négative (détérioration intellectuelle, rechute brutale, récurrence dépressive, etc.) ou positive (amélioration thymique, rémission complète, normalisation du sommeil, stabilisation des troubles, etc.).
  - (c) **Observation médicale générale - 1014 concepts** : ces concepts regroupent des observations médicales générales qui peuvent faire référence à des symptômes tels que idées délirantes, douleur, dissociation, dépendance, décompensation, etc. Nous avons également placé dans cette branche des concepts relatifs à une pathologie ou une symptomatologie. Cette branche nécessite une organisation sémantique plus fine, avec appuie des experts.
5. **Parcours de soins - 190 concepts** : nous avons organisé dans cette branche les concepts relatifs aux parcours de soins tels que hospitalisation, antécédent médical, admission, séjour en réanimation, sortie du service, traitement, résistance au traitement, etc.
6. **Trouble - 267 concepts** : cette branche regroupe toutes les expressions de type *trouble* + *terme* (visuel, langage, neurologique, psychiatrique, obsessionnel, etc.).

#### 6.4 Les limites du module d'alignement des classifications

Lors de la présentation des deux courants de classification en section 1.3, nous avons présenté les points divergents entre l'approche catégorielle et l'approche dimensionnelle, en prenant en exemple le DSM IV TR et le DSM 5. Pour résumer notre propos, nous avons constaté que si l'un est trop rigide dans les catégories descriptives, au point de rendre difficile l'inclusion des patients, l'autre a le défaut inverse, et tend à inclure trop de personnes. Nous avons décidé d'aligner ces deux systèmes de classification opposés sur leur manière d'appréhender la description des troubles psychiatriques. La rupture qui existe entre ces deux approches perturbe donc la méthodologie que nous avons adoptée.

Le DSM IV TR s'inscrit dans une approche catégorielle. Pour entrer dans une catégorie diagnostique, le patient doit présenter un ensemble précis de symptômes. Cette démarche s'avère efficace pour la recherche, mais inadaptée à la réalité clinique. Un trouble peut se manifester de différentes manières, entre autres par le biais de symptômes comorbides qui ne sont pas pris en compte dans une approche catégorielle (par exemple, la résistance à un traitement).

Le DSM 5 vise à apporter une réponse aux limites du DSM IV TR, en se tournant vers une approche dimensionnelle. Les troubles ne sont pas définis par la dichotomie présent/absent, mais par un degré de sévérité. Les symptômes comorbides sont ainsi pris en compte dans le diagnostic. Cependant, le DSM 5 conserve un système par catégorie de troubles. C'est à l'intérieur de ces catégories que l'approche dimensionnelle apporte une plus grande flexibilité diagnostique.

La pertinence de l'alignement de ces systèmes semble importante à discuter. Nous apportons dans notre travail un alignement de la partie catégorielle de ces deux classifications. Nous alignons les troubles qui trouvent une correspondance lexicale (par le biais de l'outil ONAGUI) ou sémantique (par le biais de la CIM-10). Nous ne modélisons pas les critères diagnostiques. Dès lors, nous ne prenons pas partie dans la manière de rendre le diagnostic. Notre modèle se limite à proposer un alignement entre les troubles répertoriés dans les deux classifications. Cependant, c'est le DSM IV TR qui est aligné sur l'organisation des catégories descriptives du DSM 5. Un trouble sera donc toujours représenté selon le point de vue du DSM 5. Par exemple, le trouble de *mutisme sélectif* est dans le modèle comme dans le DSM 5, un concept classé sous le concept « Trouble anxieux », lui-même classé sous le concept « Troubles neurodéveloppementaux ». Dans le DSM IV TR, le *mutisme sélectif* est classé sous le concept « Autres troubles » lui-même classée sous le concept « Troubles première et deuxième enfance ou adolescence ». Les critères diagnostiques ne changent pas d'un manuel à l'autre. Néanmoins, dans le DSM 5 et donc dans le modèle, la notion de *trouble anxieux* s'ajoute à la description du *mutisme sélectif*. Le concept de *mutisme sélectif* du DSM IV TR est donc légèrement modifié lors de son alignement avec le DSM 5. Ce constat particulier peut s'appliquer à l'ensemble des troubles du DSM IV TR alignés sur le DSM 5. Sans une validation par des experts, l'utilisation des codes du DSM IV TR contenus dans le modèle peut donc être considérée caduque.

## 6.5 Synthèse

Dans ce chapitre, nous avons été confrontés aux difficultés liées à l'absence de modélisation des classifications psychiatriques, ainsi qu'au manque de consensus autour des catégories descriptives des troubles psychiatriques. Le codage médical ou encore l'interopérabilité des données posent des problèmes méthodologiques, dès lors qu'aucun outil ne permet de faire un lien entre les différents systèmes de classification.

Nous avons choisi de modéliser les codes de trois grandes classifications psychiatriques au sein d'un même modèle développé en OWL : le DSM 5 qui est la base du modèle, le DSM IV TR et la CIM-10 dont les codes sont alignés sur ceux du DSM 5. Le but de notre modèle est de servir de socle à un outil d'annotation des diagnostics, des codes de classifications ou encore des observations médicales. Nous ne proposons pas un outil d'aide au diagnostic, par conséquent les règles de codage des classifications psychiatriques ne sont pas modélisées. Nous avons suivi une méthode descendante pour l'alignement des classifications. Ce modèle composé des 832 codes du DSM 5, propose également 334 codes du DSM IV TR et 588 codes de la CIM-10 alignés sur ceux du DSM 5, ainsi que 290 codes alignés sur les trois classifications.

Ensuite, nous avons développé un modèle grâce à l'extraction de termes des comptes rendus d'hospitalisation de l'hôpital Sainte-Anne. Un travail important de validation des

termes, puis d'organisation conceptuelle a résulté en un modèle de près de 2500 concepts propres au domaine de la psychiatrie.

Le modèle OWL réalisé par alignement des classifications des troubles psychiatriques devra être validé par l'équipe de l'hôpital Sainte-Anne. Il reste donc à l'état de prototype. En effet, afin de garantir une utilisation adéquate du modèle, une validation des alignements réalisés en extension aux recommandations des classifications est indispensable. Nous préconisons que cette validation soit faite en suivant le processus de validation que nous avons expérimenté, pour le premier module développé et que nous présentons dans le chapitre suivant, qui clôture ce travail de thèse. En parallèle, toute l'articulation entre les concepts issus des classifications et ceux issus des CRH reste à faire, en interaction avec les acteurs du domaine.



# La validation de l'ontologie sur les facteurs sociaux et environnementaux avec la méthode interactive LOVMI

## Sommaire

---

<b>7.1 Validation de la structure de l'ontologie sur les facteurs sociaux et environnementaux</b>	<b>134</b>
7.1.1 La validation de la consistance - étape 1	134
7.1.2 La validation de la structure avec OOPS! - étape 2	134
7.1.3 La validation des labels à l'aide de requêtes SPARQL - étape 3	137
7.1.4 La validation du choix des labels - étape 4	139
<b>7.2 Validation sémantique de l'ontologie sur les facteurs sociaux et environnements</b>	<b>139</b>
7.2.1 La validation par interaction avec des acteurs du domaine modélisé - étape 5	139
7.2.2 La validation par intégration du modèle dans une application sémantique - étape 6	143
<b>7.3 Proposition de la méthode LOVMI pour la validation d'ontologies</b>	<b>145</b>
<b>7.4 Synthèse</b>	<b>148</b>

---

*Nous arrivons dans ce chapitre à la dernière étape de la création du modèle ontologique : sa validation. Alors que la validation de la structure des ontologies peut être aisément réalisée à l'aide de méthodes semi-automatiques, la validation de la sémantique nécessite l'expertise de personnes compétentes dans le domaine modélisé. Les acteurs du domaine demeurent les possesseurs de la connaissance encyclopédique et pratique qui peut faire défaut à l'ontologue. Cependant, ces acteurs ne disposent pas forcément de compétences en développement ontologique. Dans le cas de notre projet, nous avons également à faire face au manque de disponibilité des acteurs, pour réaliser cette tâche de validation. Nous nous sommes alors posé les questions suivantes : quels outils nous permettent de valider la structure de notre ontologie ? Comment organiser la validation sémantique avec les acteurs du domaine, en tenant compte de leur inexpérience en ingénierie des connaissances ? Et comment rendre possible une validation sémantique de l'ontologie efficiente et rapide, sans perdre en qualité ? En partant de notre état de l'art et de notre problématique personnelle, nous avons cherché à définir un cadre méthodologique pour la validation d'une ontologie : cette méthode se base (1) sur des critères de validation issues de la littérature, (2) un chaînage d'outils pour réaliser une validation semi-automatique de la structure de l'ontologie, et (3) des entretiens semi-directifs réalisés en interaction avec les experts du domaine pour valider la sémantique, l'aspect social du modèle. Dans ce chapitre, nous présentons nos travaux qui ont menés à la proposition de la méthode LOVMI pour LES ONTOLOGIES VALIDÉES PAR MÉTHODE INTERACTIVE. Nous avons expérimenté cette méthode sur le module des facteurs sociaux et environnementaux d'ONTOPSYCHIA (ONTOPSYCHIAFSE). Dans une première section, nous décrivons le processus de validation de la structure. Une deuxième section est consacrée à la validation de la sémantique en interaction avec les acteurs du domaine. Enfin, une troisième section résume la totalité de notre méthode.*

## **7.1 Validation de la structure de l'ontologie sur les facteurs sociaux et environnementaux**

La validation de la structure s'apparente à la validation formelle du modèle, sans tenir compte des connaissances modélisées (voir section 4.2).

### **7.1.1 La validation de la consistance - étape 1**

Nous avons fait le choix d'utiliser le raisonneur HERMIT pour valider la consistance d'ONTOPSYCHIA. Ce raisonneur est intégré à l'éditeur d'ontologies Protégé 5, que nous utilisons pour le développement de nos modèles. HERMIT nous a ainsi permis de vérifier que notre ontologie ne contenait pas de classes contradictoires et ce au fur et à mesure de la construction d'ONTOPSYCHIA.

### **7.1.2 La validation de la structure avec OOPS ! - étape 2**

Suite à notre étude de l'état de l'art sur la validation de la structure d'une ontologie, nous avons choisi d'utiliser l'outil OOPS ! (présenté en section 4.2.2) pour la validation de la structure, et ce pour plusieurs raisons : disponibilité de l'outil (l'utilisation ne requiert

aucune installation préalable), mises à jour régulières<sup>1</sup>, critères utilisés qui ont été définis suite à une étude du domaine (voir 4.1.4), possibilité de conserver son code source privé, gratuité, indépendance du module, utilisation sous n'importe quel navigateur web.

L'application prend en entrée une ontologie et le catalogue de *pitfalls*, dans le but de produire en sortie un résultat d'évaluation. Elle utilise les technologies suivantes : Java EE, HTML, JQuery, JSP et CSS. L'interface utilisateur Web permet d'entrer l'URI pointant vers l'ontologie ou le document RDF décrivant l'ontologie à analyser.

Le catalogue de *pitfalls* répertorie chacun des *pitfalls* identifiés comme tels avec une description en langage naturel et son implémentation en langage Java. Actuellement, le catalogue répertorie 41 *pitfalls* en langage naturel, dont 33 sont implémentés en Java et détectable automatiquement. Le contenu du catalogue est consultable à l'adresse suivante : <http://oops.linkeddata.es/catalogue.jsp>.

Une fois l'ontologie analysée, OOPS! renvoie une page avec les *pitfalls* détectés. En cliquant sur le nom du *pitfalls* identifié, il est possible d'avoir le détail de l'analyse accompagné des explications relatives au *pitfalls*.

### Résultats sur le module facteurs sociaux et environnementaux

L'analyse de notre ontologie (composée au moment de cette analyse de 1 484 concepts et 253 relations) a permis d'identifier 8 *pitfalls* (voir figure 7.1). Nous les avons répertoriés dans la Table 7.2 et nous avons pu en déduire les erreurs suivantes :

- (1) des problèmes d'import/export sous Protégé qui ne rattache pas toujours les concepts à la racine de l'ontologie mère ;
- (2) OOPS! nous signale que nous avons défini de mauvaises relations inverses. Cependant, l'outil fait ici une erreur, car ces deux relations sont effectivement inverses. La détection de cette erreur par OOPS! provient en fait d'une mauvaise définition des domaines et co-domaines.
- (3) OOPS! n'évalue que le modèle RDF et ne prend donc pas en charge le vocabulaire Simple Knowledge Organization System (SKOS), nous ne pouvons pas évaluer les `prefLabel` et `altLabel` – ce point a donc été réalisé à l'aide de requêtes en SPARQL Protocol and RDF Query Language (SPARQL) (présenté en section 7.1.3) ;
- (4) presque la moitié de nos propriétés n'étaient pas complètement définies en termes de domaine ou/et co-domaine ;
- (5) indiquent des erreurs dans les relations inverses ;
- (6) certains co-domaines et domaines ont été définis par l'intersection de classes au lieu d'être définis par l'union de ces classes ;
- (7) plusieurs conventions de nommage (typographies) ont été utilisées au sein de l'ontologie (ex : « Enseignement\_Secondaire » versus « PostCure ») ;
- (8) l'ontologie n'est associée à aucune licence.

Cette analyse nous a particulièrement aidé dans la validation de nos propriétés. En effet, l'absence de définition des domaines et co-domaines étaient bien souvent dû à l'inutilité de la relation dans notre modélisation. C'est pourquoi le modèle finale a 72 relations de moins que le modèle initial.

---

1. <http://oops-ws.oeg-upm.net/>



## Evaluation results

It is obvious that not all the pitfalls are equally important; their impact in the ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

- **Critical** 🚫 : It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.
- **Important** ⚠️ : Though not critical for ontology function, it is important to correct this type of pitfall.
- **Minor** 🟡 : It is not really a problem, but by correcting it we will make the ontology nicer.

[Expand All] | [Collapse All]

<b>Results for P04: Creating unconnected ontology elements.</b>	<b>2 cases   Minor</b> 🟡
<p>Ontology elements (classes, object properties and datatype properties) are created isolated, with no relation to the rest of the ontology.</p> <ul style="list-style-type: none"> <li>This pitfall appears in the following elements: <ul style="list-style-type: none"> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#Concept_Vie_Sociale">http://www.limics.fr/OntoPsychia_VieSociale#Concept_Vie_Sociale</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#Etre_Vivant">http://www.limics.fr/OntoPsychia_VieSociale#Etre_Vivant</a></li> </ul> </li> </ul>	
<b>Results for P05: Defining wrong inverse relationships.</b>	<b>2 cases   Critical</b> 🚫
<p>Two relationships are defined as inverse relations when they are not necessarily inverse.</p> <ul style="list-style-type: none"> <li>This pitfall appears in the following elements: <ul style="list-style-type: none"> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#a_pour_curateur">http://www.limics.fr/OntoPsychia_VieSociale#a_pour_curateur</a> may not be inverse of <a href="http://www.limics.fr/OntoPsychia_VieSociale#est_curateur_de">http://www.limics.fr/OntoPsychia_VieSociale#est_curateur_de</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#a_partie">http://www.limics.fr/OntoPsychia_VieSociale#a_partie</a> may not be inverse of <a href="http://www.limics.fr/OntoPsychia_VieSociale#est_partie_de">http://www.limics.fr/OntoPsychia_VieSociale#est_partie_de</a></li> </ul> </li> </ul>	
<b>Results for P08: Missing annotations.</b>	<b>1736 cases   Minor</b> 🟡
<b>Results for P11: Missing domain or range in properties.</b>	<b>153 cases   Important</b> ⚠️
<b>Results for P13: Inverse relationships not explicitly declared.</b>	<b>241 cases   Minor</b> 🟡
<b>Results for P19: Defining multiple domains or ranges in properties.</b>	<b>3 cases   Critical</b> 🚫
<b>Results for P22: Using different naming conventions in the ontology.</b>	<b>ontology*   Minor</b> 🟡
<b>Results for P41: No license declared.</b>	<b>ontology*   Important</b> ⚠️
<b>SUGGESTION: symmetric or transitive object properties.</b>	<b>5 cases</b>
<p>The domain and range axioms are equal for each of the following object properties. Could they be symmetric or transitive?</p> <ul style="list-style-type: none"> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#peut_avoir_pour_consequence">http://www.limics.fr/OntoPsychia_VieSociale#peut_avoir_pour_consequence</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#a_partie">http://www.limics.fr/OntoPsychia_VieSociale#a_partie</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#est_relatif_a">http://www.limics.fr/OntoPsychia_VieSociale#est_relatif_a</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#a_pour_cause">http://www.limics.fr/OntoPsychia_VieSociale#a_pour_cause</a></li> <li>› <a href="http://www.limics.fr/OntoPsychia_VieSociale#a_pour_consequence">http://www.limics.fr/OntoPsychia_VieSociale#a_pour_consequence</a></li> </ul>	

FIGURE 7.1 – Résultats de l'évaluation de notre module *facteurs sociaux et environnementaux des maladies psychiatriques* par OOPS!.

## 7.1 Validation de la structure de l'ontologie sur les facteurs sociaux et environnementaux

TABLEAU 7.1 – Résultats de l'analyse de OOPS! sur le module *facteurs sociaux et environnementaux des maladies psychiatriques*.

Numéro et nom du <i>pitfall</i>	Catégorie(s) du <i>pitfall</i>	Nombre de cas/ Ontologie totale	Importance :
(1) P04 : Création d'éléments non-connectés à l'ontologie	Problèmes de modélisation	2	Mineure
(2) P05 : Mauvaise définition d'une relation inverse	Problèmes de modélisation	2	Critique
(3) P08 : Annotations manquantes	Compréhension humaine	1736	Mineure
(4) P11 : Domaine et co-domaine des propriétés non-défini	Compréhension humaine Problèmes de modélisation	153	Importante
(5) P13 : Relation inverse non-définie	Compréhension humaine Problèmes de modélisation	241	Mineure
(6) P19 : Définition multiple de domaine et de co-domaine	Compréhension humaine Consistance logique Problèmes de modélisation	3	Critique
(7) P22 : Utilisation de différentes conventions de nommage	Compréhension humaine	ontologie	Mineure
(8) P41 : Pas de licence déclarée	Compréhension humaine	ontologie	Important

### Résultats sur le module maladies psychiatriques

L'analyse de OOPS! sur la structure de notre module sur l'alignement des classifications, qui regroupe des concepts du DSM IV TR, du DSM 5 et de la CIM-10 a détecté quatre *pitfalls*.

TABLEAU 7.2 – Résultats de l'analyse de OOPS! sur le module des *trouble psychiatriques*.

Numéro et nom du Pitfall	Catégorie(s) du pitfall	Nombre de cas	Importance :
(1) P04 : Création d'éléments non-connectés à l'ontologie	Problèmes de modélisation	3	Mineure
(2) P08 : Annotation manquante	Compréhension humaine	832	Mineure
(3) P22 : Utilisation de différentes conventions de nommage	Compréhension humaine	ontologie*	Mineure
(4) P41 : Pas de licence déclarée	Problèmes de modélisation	ontologie	Importante

### 7.1.3 La validation des labels à l'aide de requêtes SPARQL - étape 3

Cette étape est essentielle à la validation de notre ontologie, car notre but est d'utiliser ONTOPSYCHIA pour l'annotation de texte libre. Nous devons donc être certain que la terminologie associée à notre ontologie est complète et en adéquation avec le domaine modélisé. PROTÉGÉ ne propose pas de module pour la vérification des annotations, en particulier celles concernant les labels préférentiels et alternatifs. Comme nous venons de le voir, OOPS! propose une vérification des labels, mais uniquement pour le langage

RDF. Afin de vérifier nos labels, nous avons donc utilisé le langage SPARQL, qui permet de faire des requêtes sur le vocabulaire SKOS (illustré figure 7.2).

La validation du label implique de vérifier que tous les concepts possèdent un label préférentiel unique en anglais et un label préférentiel unique en français, selon les recommandations de SKOS présentées en section 2.2.3. Nous utilisons des requêtes SPARQL pour extraire les concepts qui ne correspondent pas à ces critères. Ensuite, nous sommes en mesure de leur fournir un label préférentiel ou de supprimer un ou plusieurs labels préférentiels, au cas où plusieurs labels préférentiels auraient été définis avec la même langue, et pour le même concept.

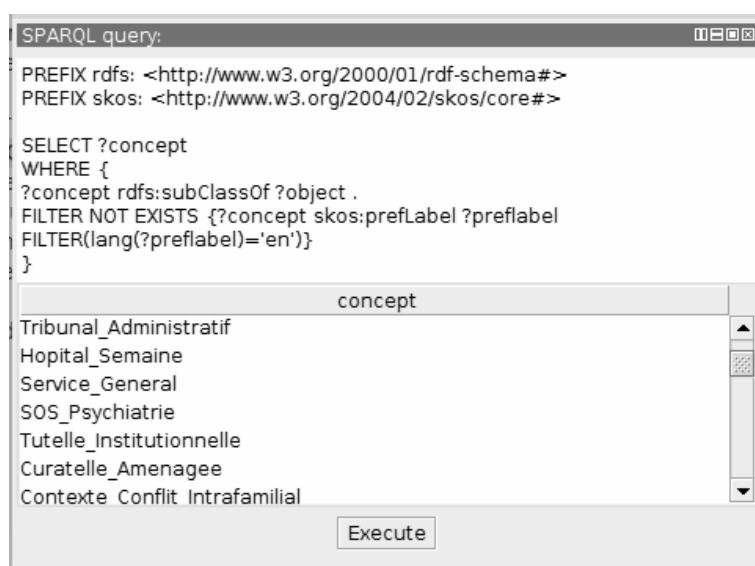


FIGURE 7.2 – Résultat de la requête (sous Protégé) permettant d'extraire les concepts sans PrefLabel en anglais.

Lors de la validation du module facteurs sociaux et environnementaux, nous avons compté trois labels préférentiels manquants pour le français et 186 labels préférentiels manquants pour l'anglais. Ce manque en anglais s'explique par un nombre important de concepts juridiques et éducatifs, qui ne peuvent pas tous être traduits du français à l'anglais. Ces notions couvrent la loi française ou le système éducatif français, et non un système international qui pourrait être équivalent pour tout le monde. Pourtant, pour être cohérent avec nos choix de modélisation, nous avons décidé de fournir une traduction lexicale (donc un label préférentiel en anglais) à un maximum d'entre eux. Lorsqu'il était absolument impossible de trouver une traduction lexicale appropriée, nous leur avons attribué le label préférentiel « No \_Equal \_Word ». À ce jour, nous comptons 38 de ces concepts sans label préférentiel approprié en anglais.

Lors de la validation des labels, nous avons également trouvé un total d'environ dix labels préférentiels en anglais, et deux labels préférentiels en français qui n'étaient pas uniques. Enfin, nous avons compté environ 35 labels préférentiels semblables au niveau graphique, mais non attribués à la même langue. Par exemple, le mot « discrimination » a la même graphie en français et en anglais, donc ce mot était deux fois en label préférentiel : une fois avec le tag anglais et une fois avec le tag français.

### 7.1.4 La validation du choix des labels - étape 4

Le choix des labels préférentiels et alternatifs a été réalisé à l'aide de la méthode développée par [Aimé et Charlet \[2012\]](#). Cette méthode s'appuie sur la biomimétique cognitive. Les auteurs précisent que chaque individu ou groupe d'individus se représentent différemment les termes associés à un concept. Leur méthode vise donc à l'évaluation du gradient de prototypicalité lexical, pour chaque terme de l'ontologie. Le but est de pouvoir déterminer le meilleur label préférentiel associé à un concept, en adéquation avec l'usage en contexte et dans un temps donné.

L'évaluation est réalisée avec une ontologie de domaine, et un corpus de textes représentant le domaine de l'ontologie. Les auteurs se basent ensuite sur le calcul de la saillance des termes, dépendant de deux mesures : (1) le calcul du poids selon la position du terme dans la structure d'un document ; (2) le calcul du poids selon la nature du document dans lequel apparaît le terme. Plus un terme est saillant, plus il est considéré proche du concept.

Le résultat de l'application de cette méthode nous a permis de découvrir que pour 76 concepts (environ 5% des concepts), notre choix de label préférentiel ne correspondait pas à l'utilisation du concept dans le corpus. Par exemple, certains acronymes sont plus courants que leur forme développée, telle que SAMU pour « service d'aide médicale urgente ».

## 7.2 Validation sémantique de l'ontologie sur les facteurs sociaux et environnementaux

La validation de la sémantique s'apparente à la validation de la sémantique des connaissances. Il s'agit de valider que le modèle conceptuel est en adéquation avec la réalité qu'il représente (voir section 4.3). Nous présentons ici une méthode pour réaliser des entretiens semi-directifs avec des acteurs du domaine modélisé.

### 7.2.1 La validation par interaction avec des acteurs du domaine modélisé - étape 5

La méthode interactive de validation que nous avons mis en place avec les acteurs du domaine a pour but d'établir un consensus sur la modélisation. Elle repose sur une communication entre les acteurs et les ontologues ayant participé au développement d'ONTOPSYCHIA.

#### Préparation de la rencontre avec les acteurs du domaine

Pour intéresser les acteurs du domaine, nous avons communiqué sur notre projet par le biais de réunions, présentations des outils ou présentations des retombées. Suite à cela, nous avons mis notre ontologie à disposition des acteurs du domaine via WEB-PROTÉGÉ. Chaque acteur pouvait visualiser l'arborescence conceptuelle de l'ontologie (incluant axiomes et relations). Nous leur avons ensuite proposé de se retrouver par petits groupes, pour leur permettre de discuter ensemble de manière interactive de la modélisation, et d'apporter leurs critiques et résolutions en cas de désaccord.

### Déroulement d'une rencontre

Les séances de validation se sont déroulées par groupes de deux (un psychologue clinicien et un psychiatre) et ont duré environ deux heures. Chaque acteur disposait de son propre ordinateur et d'un accès à l'ontologie mise en ligne sur WEBPROTÉGÉ. Chaque groupe devait travailler en interaction sur les mêmes concepts, afin de lancer des discussions et débats. Les conversations étaient enregistrées pour permettre à l'ontologue de conserver une trace de la totalité des entretiens. Les acteurs étaient invités à laisser un commentaire en texte libre sur WEBPROTÉGÉ, sous la forme d'un résumé des points abordés au cours des discussions, sur un concept ou une branche de concepts. Cela contribuait à l'interaction entre les acteurs. Ceux non présents durant la séance de validation pouvaient avoir accès aux discussions, et y répondre ou participer. Une fois ces recommandations posées, l'ontologue n'a pas donné plus d'indications aux acteurs. Il interagissait avec eux uniquement pour expliciter des choix de modélisation jugés ambigus par les acteurs. Par exemple, le concept « éducation » peut référer à « éducation scolaire » ou « éducation parentale », dans ce cas, il est important de désambiguïser verbalement la portée sémantique du concept (même si la définition du concept est déjà incluse dans l'ontologie.)

### Résultats

Pour valider le module sur les facteurs sociaux et environnementaux qui peuvent affecter la vie du patient, nous avons organisé quatre séances de validation. Chaque séance a été réalisée avec l'ontologue qui a développé le modèle et par au moins deux praticiens (un psychologue clinicien et un psychiatre). Ces séances nous ont permis de valider intégralement la taxonomie des classes. Nous avons calculé que nous pouvons valider en moyenne 164 concepts par heure. Cependant, chaque concept ou branche de concepts ne nécessite pas le même temps de validation. Dans notre modélisation, nous avons constaté que les concepts se répartissent en deux catégories selon le degré de subjectivité indiqué. Plus un concept est subjectif, plus la validation est longue. Le temps spécifié dans les exemples suivants est basé sur les enregistrements de l'entretien du premier groupe, de la première phase de validation. En outre, les séances de validation ont permis un important enrichissement conceptuel.

**Les concepts peu soumis à interprétation :** concernent les branches qui répertorient des concepts dont la signification est stable dans notre domaine de connaissances. Nous pouvons citer par exemple les branches qui modélisent « l'éducation scolaire », tel « établissement scolaire » ou « formation scolaire ». Nous comptons 124 de ces concepts. Ils ont été validés très rapidement, en survolant la hiérarchie. Nous avons enregistré 43 secondes de conversation et cinq commentaires écrits pour l'ensemble de ces concepts.

**Les concepts soumis à une interprétation définitoire, contextuelle ou personnelle :** nous devons les définir avec les acteurs du domaine. Par exemple, un « compagnon » est-il perçu différemment d'un « mari » ? Et donc utilisé différemment dans leur langage du domaine (interprétation définitoire). Ou encore le terme « relation intime » indique-t-il dans leur contexte, leur référentiel, une « relation affective très proche entre deux personnes » ou

## 7.2 Validation sémantique de l'ontologie sur les facteurs sociaux et environnementaux

TABLEAU 7.3 – ONTOPSYCHIA FSE avant la première phase de validation avec les acteurs.

Nom de la branche	Nombre de concepts	Numéro de l'équipe d'acteurs	Nombre de commentaires écrits
Concept lié à un changement	14	Équipe 1	7
Concept à connotation négative	115	Équipe 2	26
Concept lié à la justice et la loi française	3	Équipe 1	1
Concept lié à l'éducation	124	Équipe 1	7
Concept lié à un événement	13	Équipe 1	14
Concept lié à un lieu	80	Équipe 1	9
Concept lié à une situation économique	2	Équipe 2	0
Concept lié au médical	4	Équipe 2	1
Concept lié à la religion	2	Équipe 2	0
Concept lié au relationnel	152	Équipe 2	26
Concept lié à un processus	149	Non validé	NC
Total	509	2 équipes de 2 personnes à chaque séance	91

une « relation sexuelle » ? (interprétation contextuelle). Dans le cas d'interprétation définitive et contextuelle, nous avons compté 97 concepts validés en à peu près 40 minutes. Cela correspond à un temps de validation dans la moyenne. Enfin, le sens de certains concepts peut être perçu très différemment d'un individu à un autre (interprétation personnelle). Par exemple, le concept « licenciement » est une « rupture du contrat de travail », mais peut être ressenti de façon négative ou au contraire de façon positive – soulagement – dans le cas d'une personne traversant un burn-out. Ces doubles interprétations peuvent entraîner des modélisations incorrectes, si elles ne sont pas discutées avec les acteurs. Ces concepts ont amené des conversations plus denses, pour qu'un consensus autour de leur modélisation s'établisse au sein du groupe. Nous avons compté 13 de ces concepts, qui ont entraîné environ 15 minutes de discussions et généré sept commentaires (soit plus d'une minute de validation pour chaque concept).

Nous avons pu tirer plusieurs constats de ces séances de validation. Dans ces groupes de deux acteurs du domaine, aucun dominant n'est apparu. Chacun intervenait selon son expérience et ses compétences professionnelles. La visualisation totale de la hiérarchie conceptuelle les a aidés à comprendre le sens des concepts, et donc leur importance ou non dans l'ontologie. Elle a mis en avant les stratégies de modélisation et facilité la validation de branches entières de concepts, comme ce fut le cas pour les concepts liés à l'éducation scolaire. La visualisation des axiomes qui définissent les classes a également permis aux acteurs de constater des manques. Par exemple, pour l'ontologue, les concepts « maison » et « foyer » sont des lieux d'habitation, pour les acteurs il était essentiel d'ajouter la définition du type de logement en tant que « collectif » ou « individuel ». Ainsi, ce qui pourrait passer pour une faiblesse dans notre système, ou une difficulté pour les acteurs, nous a permis un gain de temps non négligeable lors de la validation.

Dans le tableau 7.3, nous fournissons les données générées par la première phase de

TABLEAU 7.4 – ONTOPSYCHIA FSE après la première phase de validation avec les acteurs.

Nom de la branche	Nombre de concepts	Observations
Concept lié à un changement	34	Enrichissement selon les remarques des acteurs
Concept lié à la justice et la loi française	199	Enrichissement selon les demandes des acteurs. Nous avons repris l'information disponible sur les sites internet du gouvernement français et du Ministère de la Justice.
Concept lié à l'éducation	132	Enrichissement selon les demandes des acteurs (e.g. conseil de classe ou conseil disciplinaire)
Concept lié à un événement	216	Enrichissement selon les remarques des acteurs. Nous avons repris en particulier les concepts des branches « Concept lié à un processus » et « Concept à connotation négative ».
Concept lié à un lieu	108	Enrichissement selon les remarques des acteurs, spécifiquement avec des concepts relatifs aux lieux médicaux (e.g. clinique, maternité, service). Nous avons repris l'information disponible sur les sites internet du gouvernement français et du Ministère de la Santé.
Concept lié au relationnel	120	Re-organisation selon les remarques des acteurs.
Total	809	La majorité des changements concernent un enrichissement avec des nouveaux concepts ou résultent d'une fusion entre différentes branches du modèle

validation. Nous comptons 509 concepts validés par deux équipes composées chacune d'un psychiatre et d'un psychologue clinicien. L'équipe 1 a validé 234 concepts et l'équipe 2 a validé 275 concepts. Ainsi, la validation se situe en moyenne à 127 concepts validés par heure. La majorité des commentaires ont été faits pour demander (a) un enrichissement de la branche, (b) un déplacement d'un concept à une autre branche ou (c) une fusion entre différentes branches. Certaines branches n'avaient pas de commentaires écrits en raison de leur faible nombre de concepts (par exemple deux pour les branches liées à la religion). Après cette première validation, nous avons opéré plusieurs changements pour se conformer aux recommandations des acteurs. Nous observons ces changements dans le tableau 7.4, accompagnés de quelques observations.

Dans le tableau 7.5, nous fournissons des données sur la seconde phase de validation. Elle nous a permis de valider les concepts qui avaient été ignorés par manque de temps, lors de la première phase de validation. Deux équipes sont intervenues. Comme lors des premières séances de validation, chaque équipe était composée d'un psychiatre et d'un psychologue clinicien. Ces équipes ont validé quatre des branches principales de la catégorie des concepts « vie sociale », ainsi que la totalité des concepts des deux autres catégories « être humain » et « groupe ». Les équipes ont validé une moyenne de 201 concepts

## 7.2 Validation sémantique de l'ontologie sur les facteurs sociaux et environnementaux

TABEAU 7.5 – La taxonomie d'ONTOPSYCHIA FSE durant la deuxième phase de validation avec les acteurs.

Nom de la branche	Nombre de concepts	Numéro de l'équipe d'acteurs	Nombre de commentaires écrits
Concept lié au comportement	48	Équipe 3	3
Concept lié aux conditions de vie	62	Équipe 4	23
Concept de caractéristique	68	Équipe 3	22
Concept lié au sentiment	63	Équipe 3	31
Concept de la branche principale « Groupe »	135	Équipe 3	27
Concept de la branche principale « Être Vivant »	350	Équipe 4	9
Total	805	2 équipes de 2 personnes à chaque séance	115

par heure, soit plus que lors de la première phase de validation. Cela s'explique par l'impact des remarques faites lors de la première phase de validation par les acteurs. En effet, ces remarques ont été utilisées pour améliorer la conceptualisation des branches qui n'avaient pas encore été validées. Cette augmentation de l'efficacité des entretiens de validation s'expliquent également par les contraintes professionnelles des acteurs, au moment de la deuxième phase de validation. En effet, ces derniers disposaient de moins de temps à consacrer à la validation.

Bien que les experts n'aient pas été facilement disponibles pour la validation, leur participation pendant les entretiens a été pertinente et productive. Ils se sont impliqués et intéressés au projet, ainsi qu'à la tâche qui leur était demandée. À l'inverse, lorsqu'on leur a demandé de procéder à la même tâche de validation, mais seul, sans interaction avec l'ontologue ni avec leurs collègues, ils n'ont pas montré le même niveau d'implication.

### 7.2.2 La validation par intégration du modèle dans une application sémantique - étape 6

La dernière étape du processus de validation consiste à utiliser l'ontologie dans une application dédiée. ONTOPSYCHIA a pour finalité d'être intégrée à des outils d'annotation de documents. Nous avons donc conçu un prototype de système d'annotation, pour trouver automatiquement des concepts dans les CRH. Le but de cette étape est de compléter la validation du modèle.

#### Présentation d'Unitex

Nous avons développé notre système avec Unitex / GramLab<sup>2</sup>. Unitex est un système multilingue de traitement de corpus en langage naturel, basé sur la technologie des automates d'états finis. Cet outil permet de gérer des ressources électroniques telles que des dictionnaires ou des grammaires et de les appliquer à des textes. Outre les importantes capacités d'analyse de ce logiciel, nous avons aussi opté pour cet outil du fait de sa libre

2. <http://www-igm.univ-mlv.fr/~unitex/index.php?page=0>



distribution sous les termes d'une licence LGPL<sup>3</sup>. Unitex est principalement développé par Sébastien Paumier à l'Institut Gaspard-Monge (IGM), Université de Paris-Est Marne-la-Vallée (France). Le concept de ce logiciel est né à LADL (Laboratoire d'Automatique Documentaire et Linguistique), sous la direction de son directeur, Maurice Gross.

## Méthode

Afin d'annoter notre corpus avec les concepts d'ONTOPSYCHIA FSE, nous avons élaboré des dictionnaires, à partir de la taxonomie de classes de l'ontologie. Un dictionnaire est composé d'entrées lexicales accompagnées d'informations grammaticales, sémantiques et/ou flexionnelles. Dans notre cas, nous indiquons uniquement le concept parent de l'entrée lexicale. Dans l'exemple suivant, « menace de mort » est l'entrée lexicale qui a pour concept parent « Menace ».

```
menace auto-agressive,.Menace
menace de suicide,.Menace auto-agressive
menace de défenestration,.Menace
menace hetero-agressive,.Menace
menace de licenciement,.Menace
menace de mort,.Menace
```

Nous avons construit autant de dictionnaires qu'ONTOPSYCHIA FSE contient de classes principales sous la classe majeure « Concept de la vie sociale » soit un total de dix dictionnaires. Nous avons ajouté un dictionnaire pour la branche des « êtres humains » et un autre pour la branche des « groupes ». Nous avons inclus les labels préférentiels et alternatifs.

Après avoir appliqué ces dictionnaires sur notre corpus, Unitex repère les entrées lexicales dans le texte. Ensuite, pour annoter ces entrées dans le texte, il est nécessaire d'utiliser des graphes de dictionnaires. Nous appelons donc le dictionnaire dans un graphe pour trouver le mot en contexte, et ajouter une balise sémantique autour de ce mot. L'entrée ressemblera à :

```
<Menace>menace auto-agressive</Menace>
<Menace>menace de défenestration</Menace>
<Menace>menace hetero-agressive</Menace>
<Menace>menace de licenciement</Menace>
<Menace>menace de mort</Menace>
```

Nous avons testé notre système sur un échantillon de 20 CRH, soit environ 15 270 mots. Nous avons envisagé la création d'un gold standard (annotation manuelle du corpus), afin de comparer ensuite les résultats obtenus par le système. Cependant, nous avons rencontré des difficultés à identifier les concepts sociaux, lors de la lecture des CRH. De plus, un seul annotateur était disponible pour réaliser ce travail, qui s'avérait très prenant en terme de temps. Par conséquent, nous avons décidé de changer notre méthode et de construire un silver standard. Nous avons commencé par annoter les CRH avec le système, ensuite nous avons examiné les annotations, pour décider de leur validité ou non et identifier les annotations manquantes.

---

3. Cette licence permet de réutiliser librement le code source du logiciel. Elle permet également d'inclure du code source dans d'autres logiciels, y compris des logiciels non libres.

### 7.3 Proposition de la méthode LOVMI pour la validation d'ontologies

---

#### Résultats

TABLEAU 7.6 – Résultats de l'annotation des CRH avec ONTOPSYCHIA.

Concepts...	Nombre
...annotés	724
...vrais positifs - VP (annotations correctes)	607
...faux positifs - FP (annotations incorrectes)	117
...faux négatifs - FN (annotations manquantes)	134
Précision ( $VP / (VP+FP)$ )	83.8%
Rappel ( $VP / (VP+FN)$ )	81.9%
F-mesure ( $2*P*R / (P+R)$ )	82.8%

Dans le tableau 7.6, le nombre total de 724 concepts annotés comprend 706 labels préférentiels annotés et, par conséquent, 18 labels alternatifs annotés. Ce résultat confirme de nouveau que notre choix de labels préférentiels était pertinent. En outre, nous comptons 401 concepts annotés de la branche « concept de la vie sociale », 238 concepts annotés de la branche « être humain » et 67 concepts annotés de la branche « groupe ». Ensuite, nous observons un bon taux de précision et de rappel. Nous avons 83.8% d'annotations pertinentes (soit 16,2% d'annotations effectuées qui sont incorrectes) et une couverture de notre corpus de 81.9% (soit 18,1% d'annotations manquantes). Nous observons que les annotations erronées ne sont pas causées par un mauvais choix de label, mais majoritairement par une analyse erronée du contexte d'apparition du concept. Par exemple, le concept « trouble » ne se présente jamais seul dans un CRH et pourtant le système a permis son annotation sans contexte (en effet le concept trouble est toujours défini en fonction de ce qui est troublé : « trouble d'attention »). Par conséquent, ce type d'erreur indique un problème dans le système d'annotation lui-même, et non une erreur dans la modélisation du domaine. Au vu des remarques précédentes, nous pouvons déduire du test d'annotation, qu'il permet l'enrichissement lexical et conceptuel de notre ontologie. Ainsi, ce système d'annotation ou une autre application est une étape importante pour la validation et l'enrichissement des ontologies.

### 7.3 Proposition de la méthode LOVMI pour la validation d'ontologies

À partir du travail décrit jusqu'ici, nous proposons la méthode LOVMI pour la validation structurelle et sémantique d'ontologies, en six étapes (tableau 7.7).

TABLEAU 7.7 – Présentation de la méthode Lovmi.

	Type d'élément validé	Outil utilisé
Étape 1	Consistance	E1 : Raisonneur (HERMIT, PELLET, FACTPLUS)
Étape 2	Structure	E2 : OOPS !
Étape 3	Labels	E3 : Requêtes SPARQL
Étape 4	E4 : Choix du label préférentiel	Méthode développée par <b>Aimé et Charlet [2012]</b>
Étape 5	E5 : Représentation sémantique du domaine	Interaction entre ontologues et acteurs du domaine
Étape 6	E6 : Adéquation sémantique du modèle pour une application	Intégration du modèle dans une application dédiée (annotation, extraction, inférence de données)

Nous avons développé cette méthode suite à une étude de la littérature. Elle s'appuie sur les six dimensions définies par **Poveda-Villalón et al. [2012]**. Nous ne proposons pas de nouvelles techniques pour la validation d'ontologies, mais une optimisation de l'utilisation d'outils, techniques et processus déjà existants et qui ont fait leur preuve dans le domaine. Nous apportons notre expérience dans la validation de la sémantique en interaction avec les acteurs du domaine. Nous avons définis un cadre méthodologique pour la réalisation d'entretiens semi-directifs avec les acteurs du domaine. En outre, nous avons pu expérimenter notre méthodologie avec succès, pour la validation de l'ontologie sur les facteurs sociaux et environnementaux des maladies psychiatriques.

De plus, notre méthode permet d'identifier 22 (sur 24) problèmes de qualité définis par **Gherasim et al. [2012]** et présentés en annexe G. Notre méthodologie se base sur l'outil OOPS ! (étape 2) qui permet d'identifier à lui seul 14 problèmes de qualité de l'ontologie. Ensuite, le raisonneur (étape 1) permet de vérifier le *problème d'inconsistance logique* entre autre. Les requêtes SPARQL (étape 3) permettent de vérifier les *problèmes de manques d'explications textuelles* et d'*équivalences potentielles entre artefacts*. La validation du choix du label (étape 4) permet de vérifier les *problèmes d'artefacts potentiellement équivalents ou indissociables*, car il assure que les labels sont adaptés aux concepts. L'interaction avec les acteurs (étape 5) permet de vérifier aussi bien des problèmes logiques que des problèmes sociaux. Nous avons identifié cinq problèmes logiques (en particulier ceux qui concernent des *modélisations inadaptées ou inadéquates*) que les acteurs peuvent résoudre, en plus de la totalité des problèmes sociaux. Enfin, l'intégration (étape 6) dans une application dédiée permet de répondre aux problèmes liés aux *formalismes non-standards* et aux *versions inadaptées de l'ontologie*. Les deux problèmes auxquels nous n'apportons pas de réponse - *raisonnement incomplet* et *complexité de raisonnement* - sont spécifiques à la modélisation des axiomes logiques. Notre méthode se focalise sur l'interaction avec les experts, pour la validation d'ontologies contenant une complexité sémantique, dans la représentation des concepts. Nous n'apportons pas de réponse supplémentaire à ce qui est déjà proposé dans la littérature, pour la validation des axiomes logiques.

### 7.3 Proposition de la méthode LOVMI pour la validation d'ontologies

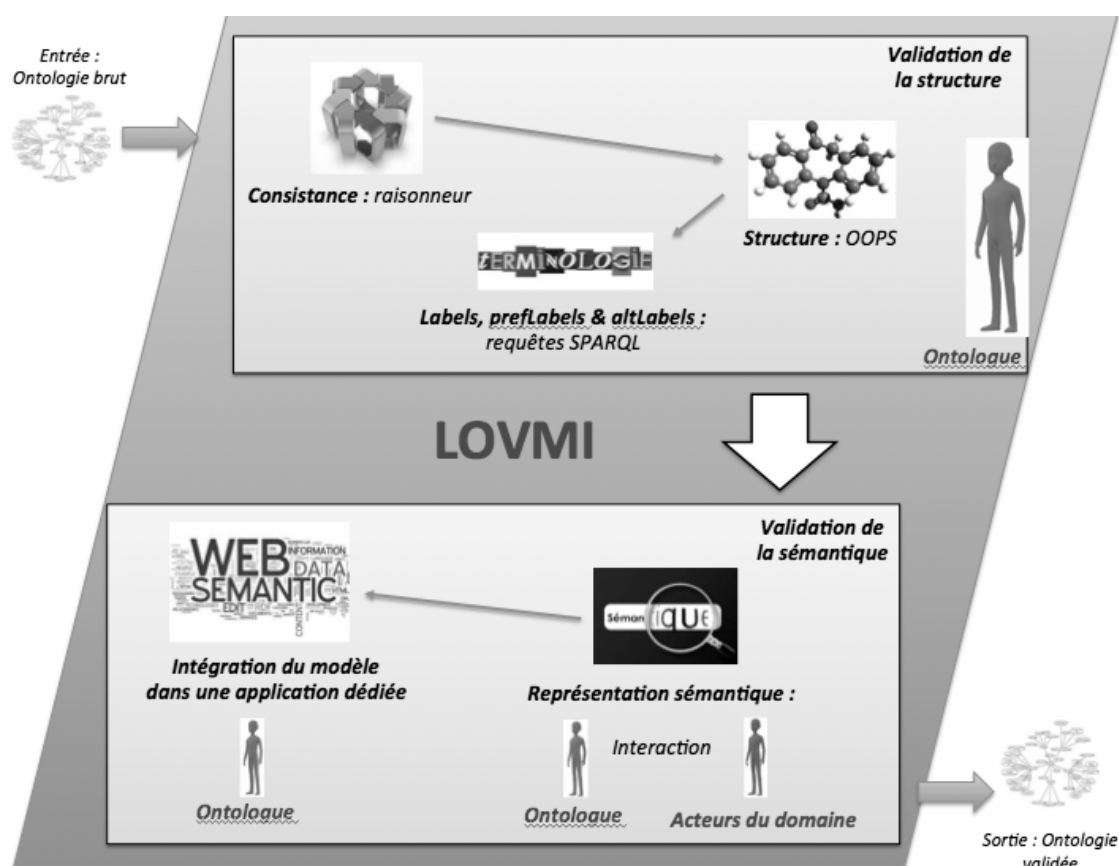


FIGURE 7.3 – Schéma modélisant le cadre méthodologique LOVMI.

La figure 7.3 illustre la méthode. Pour la validation de la structure, LOVMI utilise une série d'outils existants ayant prouvé leur efficacité dans la tâche de validation pour laquelle ils ont été développés. Pour valider la sémantique de l'ontologie, des entretiens interactifs avec des spécialistes du domaine ont été réalisés. La validation avec LOVMI s'effectue en six étapes, réalisées de manière semi-automatique et en collaboration directe avec des acteurs de terrain. Les étapes ne sont ni linéaires, ni cycliques. Elles peuvent être effectuées indépendamment les unes des autres. L'utilisateur peut sauter une étape pour y revenir plus tard. Par exemple, nous pouvons valider la hiérarchie des concepts avec les acteurs, avant de vérifier que l'ontologie contient tous les labels associés aux concepts. Cependant, il est plus adéquat de vérifier la consistance de l'ontologie avant de la faire valider par des experts. L'ordre dans lequel nous avons ordonné les étapes a donc été pensé de manière à optimiser la validation. Enfin, pour considérer qu'une ontologie a été entièrement validée, toutes les étapes doivent être complétées.

Pour mettre en avant l'impact de la validation sur l'ontologie, nous présentons dans le tableau 7.8 l'ontologie des facteurs sociaux et environnementaux avant la validation et dans le tableau 7.9 l'ontologie après la validation.

TABLEAU 7.8 – L'ontologie des « Facteurs sociaux et environnementaux des maladies psychiatriques » avant la validation avec les experts.

Module social	Classes	Relations
OntoPsychiaFSE	1478 (172 avec deux parents)	199
Concept sur la vie sociale	1039	51
Être humain	560	144 (dont 140 issues de la FHHO)
Groupe	81	2

TABLEAU 7.9 – L'ontologie des « Facteurs sociaux et environnementaux des maladies psychiatriques » après la validation avec les experts.

Module social	Classes	Relations
OntoPsychiaFSE	1345 (267 avec deux parents)	271
Concept sur la vie sociale	884	71
Être humain	560	193 (dont 140 issues de la FHHO)
Groupe	81	7

## 7.4 Synthèse

L'étude de la littérature a mis en avant l'importance de la validation de la structure des ontologies, qui peut être effectuée automatiquement à l'aide d'outils informatiques. La validation de la structure garantit l'utilisation de l'ontologie au sein d'applications dédiées, ainsi que la réutilisation à d'autres fins. Elle permet de s'assurer que les inférences sont correctes, que le contenu informatif (les méta-propriétés) peut être compris par n'importe quel ontologue, que le langage ontologique est correctement utilisé ou que les conventions sont respectées. Néanmoins, une structure correcte ne garantit pas une sémantique valide et adéquate au domaine. Le critère indiqué par [Poveda-Villalón et al. \[2012\]](#) concernant la précision de la modélisation ontologique du domaine doit donc être vérifié par des humains.

Dans ce chapitre, nous avons présenté LES ONTOLOGIES VALIDÉES PAR MÉTHODE INTERACTIVE (LOVMI). Ce travail a abouti de notre volonté de disposer d'une méthode pour vérifier la qualité de l'ontologie sous tous ses aspects, structurels et sémantique, logiques et sociaux. La méthode s'appuie sur un chaînage d'outils existants ayant fait leur preuve dans leur domaine. Nous les avons choisi selon les critères qu'ils valident. Nous avons également constaté que les méthodes actuelles ne proposent pas de support pour travailler avec les experts à la validation du modèle de connaissances. Ainsi, LOVMI propose un support méthodologique, pour réaliser des entretiens semi-directifs avec des acteurs du domaine, en vue de valider une ontologie. La méthode a été testée avec succès, pour valider le module sur les facteurs sociaux et environnementaux d'ONTOPSYCHIA.

# Conclusion

Ce manuscrit présente les travaux de recherche en informatique médicale qui ont mené au développement d'ONTOPSYCHIA, une ontologie pour la psychiatrie, composée de deux modules indépendants : (1) les facteurs sociaux et environnementaux des troubles mentaux, (2) les troubles mentaux. Notre travail a contribué à la mise en œuvre des méthodes hybrides de développement d'ontologies. En outre, nous nous sommes intéressés aux problématiques liées à la validation des ontologies et nous avons développé une méthode de validation structurelle et sémantique des ontologies, la méthode LOVMI.

La psychiatrie est une spécialité médicale qui vise à fournir un diagnostic et à traiter des troubles mentaux. Malgré des classifications internationalement reconnues, la catégorisation des patients selon des critères diagnostiques reste problématique. Les catégories actuelles peinent à prendre en compte l'hétérogénéité interindividuelle, les difficultés de délimitations des syndromes et l'influence sur les symptômes de nombreux facteurs dans l'histoire individuelle ou dans l'environnement.

Ainsi, nos travaux ont porté sur trois problématiques principales : (1) l'absence de prise en compte des facteurs sociaux et environnementaux dans la catégorisation des patients, (2) l'absence d'alignement entre les classifications actuelles, (3) l'incohérence entre la réalité clinique et les catégories descriptives des classifications.

Pour répondre à ces problématiques, nous avons développé des solutions au travers des méthodes et outils proposés en ingénierie des connaissances, et plus précisément en ingénierie des ontologies. Une ontologie informatique permet de définir un vocabulaire commun et une représentation consensuelle d'un domaine donné. Les ontologies permettent de partager de l'information aussi bien au niveau humain qu'au niveau machine. D'ailleurs, elles ont été popularisées bien avant l'essor du Web Sémantique, grâce au besoin d'intégration de données de disciplines majeures, telles que la biologie et la médecine dans les années 1990. Aujourd'hui, des disciplines habituellement laissées à l'écart des nouvelles technologies, telle que la psychiatrie, s'intéressent à leur puissance de modélisation, de partage de l'information et de raisonnement.

Notre objectif à court terme était de proposer à l'attention du SHU de Sainte-Anne, une modélisation des facteurs sociaux et environnementaux des maladies psychiatriques,

ainsi que des maladies psychiatriques elles-mêmes. La vision à long terme de nos travaux est de disposer d'un modèle pour appuyer le développement d'outils (1) pour l'analyse des facteurs de risques sociaux et environnementaux des maladies psychiatriques et (2) pour le codage des maladies et de la comorbidité. Ces outils pourront aider la prévision ou l'identification des troubles, ainsi que la prise en charge des patients au niveau médical (comme, par exemple, définir un parcours de soins mieux adapté en fonction des symptômes et des troubles) et au niveau économique (comme, par exemple, mieux anticiper les besoins financiers pour couvrir les soins).

## **Bilan des contributions proposées**

### **Sur le recueil de données**

Notre première problématique a été le choix des sources pour la construction de l'ontologie, afin de représenter le plus finement possible la connaissance des personnes du service hospitalier et son utilisation dans le contexte. Nous sommes partis sur une solution mixte : données brutes issues de comptes rendus hospitaliers (CRH) et données structurées issues de classifications utilisées au sein du service, ou reconnues dans le domaine.

Nous avons également été confrontés aux problématiques de traitement de données confidentielles, que sont les données médicales. Les corpus médicaux contiennent des informations sensibles et confidentielles. Aujourd'hui, peu d'outils permettent d'automatiser un traitement d'anonymisation rapide et efficace sur des textes en français. Notre projet a donc démarré sur une tâche de préparation de corpus minutieuse et cruciale pour la suite de nos travaux. C'est le logiciel MEDINA [Grouin *et al.*, 2009] qui nous a aidés à semi-automatiser cette tâche et à réaliser l'anonymisation de près de 8 000 comptes rendus d'hospitalisation.

### **Sur l'extraction de termes candidats**

Afin de développer notre ontologie à partir des données contenues dans les CRH, nous avons réalisé une extraction des termes candidats de notre corpus. Les extracteurs de termes disponibles librement et gratuitement pour traiter les textes en français sont là encore peu nombreux. Nous en avons testé deux qui répondaient aux exigences de notre corpus (liées en particulier à la confidentialité des données). Nous avons comparé ces deux extracteurs, afin de définir l'outil adéquat pour l'analyse de notre corpus. Les résultats de notre étude comparative montrent que l'outil BIOTEX [Lossio-Ventura *et al.*, 2014] est le plus performant pour l'extraction de termes candidats dans les corpus médicaux.

### **Sur la construction des modules ontologiques**

Nous avons fait le choix de développer une ontologie modulaire, suite à une étude de la littérature sur les méthodes de construction d'ontologies présentée au **chapitre 3**. Notre ontologie comporte à ce jour trois modules.



## Conclusion

---

Le premier module vise à modéliser les *aspects sociaux et environnementaux* qui peuvent avoir une incidence sur la vie d'une personne. L'état de l'art a montré que les thésaurus et les ontologies médicales qui modélisent des aspects sociaux dans la vie d'un patient restent encore trop peu exploités en psychiatrie. Aujourd'hui, avec le développement de la médecine de précision, nous assistons à une mise en lumière de l'approche holistique, notamment dans la prise en charge des patients en psychiatrie. Ainsi, l'analyse de la prévalence et de l'incidence des facteurs de risques sociaux et environnementaux des maladies psychiatriques semble désormais cruciale. Une meilleure compréhension de ces risques pourrait avoir des conséquences significatives aussi bien en termes de prise en charge des patients (parcours de soin et traitement par exemple) qu'en termes de politique de santé publique (coûts et durées des hospitalisations notamment). Afin de construire notre module, nous avons expérimenté les étapes de la méthodologie hybride TOREUSE2ONTO [Drame \[2014\]](#). Nous avons ainsi modélisé un peu plus de 1400 concepts et 180 relations, avec un alignement de 93 concepts sur la SNOMED-3.5VF.

Les deuxième et troisième module vise à modéliser les *troubles/maladies psychiatriques*. En première partie de la thèse, nous avons fait état d'un manque de modélisation formelle des classifications psychiatriques et observé une absence de consensus autour des catégories descriptives des troubles psychiatriques. Nous nous sommes donc interrogés sur la manière de construire une ontologie qui tiendrait compte de ces problématiques. Nous avons choisi de modéliser les codes de trois grandes classifications psychiatriques au sein d'une même ontologie : les codes du DSM 5 sont la base de l'ontologie et les codes du DSM IV TR et de la CIM-10 sont alignés sur eux. Nous avons expérimenté une méthode descendante pour l'alignement des classifications. Ce modèle composé des 832 codes du DSM 5, propose également 334 codes du DSM IV TR alignés sur ceux du DSM 5 et 588 codes de la CIM-10 alignés sur ceux du DSM 5, ainsi que 290 codes alignés sur les trois classifications. Nous avons ensuite construit un nouveau module ontologique grâce à l'extraction de termes des CRH de l'hôpital Sainte-Anne. Un travail important de validation des termes, puis d'organisation conceptuelle a résulté en un modèle de près de 2500 concepts propres au domaine de la psychiatrie. Ce module indépendant des classifications permet de compléter la modélisation des *troubles/maladies psychiatriques* en termes de description des troubles.

### Sur la validation des ontologies

En premier lieu, nous avons réalisé une étude comparative des outils de validation d'ontologies. Nous avons alors constaté que toutes ces méthodes pouvaient être divisées en deux groupes : (1) les méthodes pour valider la structure et (2) les méthodes pour valider la sémantique. Suite au constat qu'aucune méthode ne proposait la validation de ces deux axes, nous avons choisi de développer notre propre méthode de validation tenant compte à la fois de la structure et de la sémantique. Pour valider la structure de l'ontologie, son aspect logique LOVMI utilise une série d'outils existants validés scientifiquement. Pour la sémantique, l'aspect social LOVMI définit une méthodologie pour réaliser des entretiens interactifs avec des acteurs du domaine modélisé. Nous avons testé notre méthode avec succès, à l'échelle d'une ontologie de plus de 1400 concepts. Le module sur les facteurs sociaux et environnementaux d'ONTOPSYCHIA a ainsi été validé en suivant les



différentes étapes de LOVMI (résumées en section 7.3). En outre, notre méthode se veut une méthode générique qui peut être utilisée sur toutes les ontologies de domaine.

## **Limites**

### **Le module sur les facteurs sociaux et environnementaux**

La validation du module effectuée lors du test à l'échelle de la méthode LOVMI a montré de bons résultats (section 7.6). Nous considérons que notre module est utilisable et intégrable en l'état, dans des applications et outils dédiés.

Toutefois, les résultats affichent une annotation de 80% des concepts présents dans les CRH. De facto, ils montrent qu'environ 20% des concepts n'ont pas été annotés. Bien que ce chiffre soit à minorer avec les erreurs liées au système d'annotation lui-même, il montre qu'un enrichissement conceptuel (différents types de relation à ajouter par exemple) et lexical (ajout nécessaire de synonymes ou d'abréviations) du modèle est encore possible. Cet enrichissement régulier permettra d'approcher une représentation plus fine des facteurs sociaux et environnementaux décrits dans les CRH.

### **Le module sur les maladies psychiatriques**

Le module sur les maladies psychiatriques est un alignement sur le DSM 5, du DSM IV TR et de la CIM-10. Cependant, nous avons montré au cours de ce travail de recherche, que ces classifications se différencient sur bien des points et sur leur approche en particulier. Le DSM 5 s'inscrit dans une approche dimensionnelle, qui tend à décrire les troubles au travers de spectres, tels que le « spectre de troubles schizophréniques » ; le « spectre des troubles bipolaires » ou encore le « spectre autistique ». Ce concept dénote la diversité des troubles et des symptômes qui peuvent se retrouver dans une pathologie. A contrario, le DSM IV TR et la CIM-10 s'inscrivent dans une approche catégorielle, qui tend à définir des catégories précises de troubles, de syndromes décrits par un ensemble de symptômes et de faits chronologiques. Nous avons discuté plus en détail en section 6.4 l'impact sur notre ontologie de la différence d'organisation des troubles, dans ces deux approches. Dès lors, nous pouvons considérer qu'un alignement fondé (1) sur une comparaison des termes dénotant chaque concept dans les diverses classifications et (2) sur les recommandations des classifications peut s'avérer limitée voire tout simplement fausse. Par conséquent, les résultats de notre alignement doivent être discutés par des professionnels de la santé mentale. En effet, l'absence de validation sémantique de ce module ne permet pas d'en envisager son utilisation. En ce qui concerne les concepts issus des CRH, ils sont également à valider en intégralité. Ce module est donc encore en cours de développement.

### **La méthode de validation**

Le domaine médical, peut-être plus que dans d'autres domaines, pose un réel problème de disponibilité des praticiens dans les différentes phases de conception d'une ontologie qui est in fine censée représenter une vision consensuelle de leur domaine de prédilection. Aussi, afin de pallier ce manque de disponibilité, nous avons cherché à automatiser le plus grand nombre de tâches, afin d'être le moins dépendant possible des acteurs

du domaine. Or, même si nous disposons de leurs ressources, les acteurs du domaine demeurent les premiers possesseurs de la connaissance encyclopédique et pratique de leur domaine d'expertise ; ce qui n'est pas vrai pour l'ontologue – sauf si celui-ci dispose de plusieurs compétences en simultané. L'observation majeure de notre étude est donc la nécessité absolue de la participation des acteurs du domaine / des experts pour obtenir un résultat non seulement valide mais également « appropriable » par ses utilisateurs. Car ils pourront y retrouver leurs connaissances sous la forme qu'ils exploitent tous les jours. Ils ne peuvent pas être remplacés par un algorithme automatique (constat déjà effectué avec les premiers systèmes experts). En conséquence, nous recommandons, avant de planifier un projet de modélisation des connaissances, de veiller à ce que les experts, acteurs du domaine (ou toutes autres personnes compétentes dans le domaine et reconnue comme telle par ses pairs) puissent s'investir dans le projet et allouer suffisamment de temps aux activités de validation du modèle. Nous avons montré que le manque de disponibilité est une limite qui ne peut être contournée dans le cadre d'un projet de développement d'une ontologie, alors même que cette expertise des acteurs du domaine modélisé est le garant d'une ontologie de qualité.

Ainsi, les limites de notre méthode de validation LOVMI sont liées aux limites d'implication temporelle des experts, pour réaliser cette tâche. En effet, la tâche de validation la plus complexe et la plus longue est la validation sémantique, par les acteurs du domaine. Une solution pour minimiser cette contrainte a été d'utiliser des outils collaboratifs disponibles sur un navigateur web et d'organiser des sessions de validation, pour faciliter l'accès au processus de validation. Nous avons alors constaté que ce n'est pas l'intérêt des experts pour la validation qui est problématique, mais bien leur temps de disponibilité. Au travers de la validation sémantique, nous avons également souligné l'importance d'une phase de test dans une application, pour valider l'adéquation de l'ontologie, ainsi que pour enrichir le modèle conceptuel et le lexique.

En outre, notre méthode de validation ne répond pas aux questions posées sur la validation des axiomes logiques.

## Perspectives

### Un module sur les traitements

Pour compléter l'ontologie de la psychiatrie, un troisième module est en cours de développement. Ce module vise à la modélisation des *traitements médicamenteux et non médicamenteux*. Ce travail de modélisation a été réalisé dans le cadre du stage orienté recherche de [Steinberg \[2016\]](#). Pour réaliser cette ontologie, Karine Steinberg s'est appuyée sur les travaux amorcés par Xavier Aimé en 2013. L'ontologie créée de façon semi-automatique au moyen de l'outil TALEND OPEN STUDIO contenait déjà 50 000 concepts issus de sources diverses telles que : (1) la liste des Autorisation de Mise sur le Marché des médicaments (AMM) délivrée par l'Agence Nationale de Santé du Médicament et des produits de santé (ANSM) ; (2) la liste des Unité Commune de Dispensation des médicaments (flacon, ampoule, comprimé, entre autres) qui est un code commun à tous les médicaments disposant d'une AMM ; (3) la base de données ouvertes RXNorm<sup>4</sup> ; (4)

---

4. RXNorm est un catalogue de noms de médicaments standardisés, qui est intégré au méta-thésaurus Unified Medical Language System (UMLS).

la classification ATC (anatomique, thérapeutique et chimique) qui permet de classer les médicaments par groupe anatomique et thérapeutique. Cependant, cette ontologie réalisée à partir de sources de données multiples était fortement bruitée et mal structurée. L'important travail de nettoyage, formatage, traduction des sources non francophones en français ou encore enrichissement conceptuel par intégration de nouvelles bases de données a abouti à une ontologie structurée de 184 000 concepts, qui doit encore être corrigée puis testée en contexte applicatif. Ce module offrira la possibilité d'annoter les traitements dans les documents médicaux tels que les CRH. C'est une source d'information considérable, qui vient s'ajouter aux deux modules ontologiques sur les facteurs sociaux et environnementaux, et sur les maladies mentales.

### **Des outils basés sur l'ontologie**

Sur le long terme, l'ontologie a été développée afin de servir de socle à un système à base de connaissances. En effet, une fois les trois modules finis et reliés à une top ontologie, ONTOPSYCHIA pourra être intégrée dans des outils ou applications dédiés, notamment pour réaliser l'annotation conceptuelle d'un document, de la même manière que le prototype présenté en section 7.2.2. Les buts sont multiples :

1. L'analyse des facteurs sociaux et environnementaux des maladies psychiatriques aidera à en comprendre les risques et les impacts sur les troubles. Cette analyse aidera également à la prévision ou à l'identification des troubles, ainsi qu'à la prise en charge des patients. Ce premier objectif répond aux ambitions de la médecine de précision en psychiatrie.
2. L'accès à la comorbidité qui se définit par la présence simultanée de deux maladies qu'elles soient liées ou non dans les CRH répondra aux enjeux économiques PMSI<sup>5</sup>. Les CRH psychiatriques présentent, aujourd'hui, une assez forte hétérogénéité de contenu en fonction, ou non, de l'utilisation des classifications et des pratiques locales. Au delà, ces difficultés de délimitation des syndromes compliquent la définition des populations incluses dans les essais thérapeutiques et les recherches étiologiques ou de biomarqueurs. Les catégories actuelles correspondent encore à des situations hétérogènes. En termes médico-économiques, la résistance aux traitements est une caractéristique au moins aussi importante que la catégorie du diagnostic principal. Il est enfin à noter que la notion de comorbidité, donnée clinique essentielle, est souvent absente des codages de dossiers qui se limitent la plupart du temps à un item issu de la CIM-10. Par exemple, un patient présentant un syndrome schizophrénique et une insuffisance cardiaque ne pourra être codé qu'avec un diagnostic. Cela pose un réel problème dans l'indexation des dossiers, en plus des problèmes économiques dû à l'absence de la prise en compte total des pathologies. Le but de ce module est donc double, d'une part il vise à permettre d'annoter

---

5. Le Programme de médicalisation du système d'information (PMSI), vise à introduire des concepts de comptabilité analytique dans la gestion administrative des hôpitaux : les diagnostics et actes effectués dans un établissement de santé sont codés et comptabilisés, rapportés à un patient et aux différents coûts de la structure. Cela permet ainsi de bâtir des indices de coûts relatifs par groupe homogène de malades. Le PMSI utilise un système de codage international, la CIM-10, pour les diagnostics, et un système français, développé grâce à une approche ontologique : la CCAM, pour les actes. Le codage des diagnostics se fait en posant un diagnostic principal et, si nécessaire (au maximum 5), des diagnostics associés. Le PMSI a évolué vers une comptabilité qui vise à analyser le coût de chaque acte : c'est la tarification à l'activité ou T2A mais elle ne concerne pas la psychiatrie.

## **Conclusion**

---

différentes pathologies par le biais d'un identifiant unique (URI) qui permettra l'indexation des dossiers pour la constitution de cohortes et d'autre part, d'améliorer le système de codage dans le cadre des politiques économiques.

3. L'accès aux informations sémantiques contenues dans les CRH aidera à établir un profil de patient présentant une résistance aux traitements.



---

## Bibliographie

- Hempel, C. G. (1965). Fundamentals of taxonomy. *Aspects of scientific explanation*, 137 :154.
- Peirce, C. S. (1978). *Ecrits sur le signe*, volume 31. Seuil.
- Eco, U. (1984). *Semiotica e filosofia del linguaggio*. Einaudi.
- Aussenac-Gilles, N. (1989). *Conception d'une méthodologie et d'un outil d'acquisition de connaissances expertes*. PhD thesis, Toulouse 3.
- De Saussure, F. (1989). *Cours de linguistique générale : Édition critique*, volume 1. Otto Harrassowitz Verlag.
- Everaert-Desmedt, N. (1990). *Le processus interprétatif : introduction à la sémiotique de Ch. S. Peirce*. Editions Mardaga.
- WHO et al. (1992). *The ICD-10 classification of mental and behavioural disorders : clinical descriptions and diagnostic guidelines*.
- Gruber, T. R. et al. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, 5(2) :199–220. <http://www.sciencedirect.com/science/article/pii/S1042814383710083>.
- Knight, K. et Luk, S. K. (1994). Building a large-scale knowledge base for machine translation. In *AAAI*, volume 94, pages 773–778. <http://www.aaai.org/Papers/AAAI/1994/AAAI94-118.pdf>.
- Schreiber, G., Wielinga, B., de Hoog, R., Akkermans, H., et Van de Velde, W. (1994). Commonkads : A comprehensive methodology for kbs development. *IEEE expert*, 9(6) :28–37.
- Rastier, F., Cavazza, M., et Abeillé, A. (1994). Sémantique pour l'analyse.
- Gómez-Pérez, A., Juristo, N., et Pazos, J. (1995). Evaluation and assessment of knowledge sharing technology. *Towards very large knowledge bases*, pages 289–296.
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International journal of human-computer studies*, 43(5) :907–928. <http://www.sciencedirect.com/science/article/pii/S1071581985710816>.
- King, M. U. M. et Uschold, M. (1995). *Towards a Methodology for Building Ontologies*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.480.1214&rep=rep1&type=pdf>.
- Mercier, D.-J. (1996). Cryptographie classique et cryptographie publique à clé révélée. *Bulletin APMEP*, (406) :568–581. <https://hal.archives-ouvertes.fr/hal-00762806/document>.
- Prytherch, R. et Bloomberg-Rissman, J. (1996). Harrod's librarians' glossary. *Library Quarterly*, 66(3) :319–320.

- 
- Swartout, B., Patil, R., Knight, K., et Russ, T. (1996). Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, pages 138–148.
- Weibel, S. (1997). The dublin core : a simple content description model for electronic resources. *Bulletin of the American Society for Information Science and Technology*, 24(1) :9–11. <http://onlinelibrary.wiley.com/doi/10.1002/bult.70/full>.
- Fernández-López, M., Gómez-Pérez, A., et Juristo, N. (1997). Methontology : from ontological art towards ontological engineering. pages 33–40. <http://oa.upm.es/5484/>.
- Farquhar, A., Fikes, R., et Rice, J. (1997). The ontolingua server : A tool for collaborative ontology construction. *International journal of human-computer studies*, 46(6) :707–727. <http://www.sciencedirect.com/science/article/pii/S1071581996901214>.
- Guarino, N. (1997). Semantic matching : Formal ontological distinctions for information organization, extraction, and integration. In *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*, pages 139–170. Springer. [http://link.springer.com/chapter/10.1007/3-540-63438-X\\_8](http://link.springer.com/chapter/10.1007/3-540-63438-X_8).
- Charlet, J. et Bachimont, B. (1998). De l'acquisition à l'ingénierie des connaissances : applications et perspectives. *Actes des assises nationales 1998 du PRC-I3*, pages 81–84.
- Fernández-López, M., Gómez-Pérez, A., Sierra, J. P., et Sierra, A. P. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems and their applications*, 14(1) :37–46. [http://oa.upm.es/5466/1/Building\\_a\\_Chemical\\_Ontology.pdf](http://oa.upm.es/5466/1/Building_a_Chemical_Ontology.pdf).
- Biebow, B., Szulman, S., et Clément, A. J. (1999). Terminae : A linguistics-based tool for the building of a domain ontology. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 49–66. Springer. [http://link.springer.com/chapter/10.1007/3-540-48775-1\\_4](http://link.springer.com/chapter/10.1007/3-540-48775-1_4).
- Chandrasekaran, B., Josephson, J. R., et Benjamins, V. R. (1999). What are ontologies, and why do we need them? *IEEE Intelligent systems*, (1) :20–26. <http://www.ce.unipr.it/~paterli/mmSE/chandra.pdf>.
- Darmoni, S. et Joubert, M. (2000). Cismef. *Methods of information in medicine*, 39(1) :30–35. <http://www.cismef.org/CISMeF%20August%202009a.pdf>.
- Bachimont, B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, pages 305–323. [https://www.utc.fr/~bachimon/dokuwiki/\\_media/fr/ontologie-icbook.pdf](https://www.utc.fr/~bachimon/dokuwiki/_media/fr/ontologie-icbook.pdf).
- Guarino, N. et Welty, C. (2000). A formal ontology of properties. In *Knowledge Engineering and Knowledge Management Methods, Models, and Tools*, pages 97–112. [https://link.springer.com/chapter/10.1007/3-540-39967-4\\_8](https://link.springer.com/chapter/10.1007/3-540-39967-4_8).
- Schreiber, G. (2000). *Knowledge engineering and management : the CommonKADS methodology*. MIT press.

- Pinto, H. S. et Martins, J. (2000). Reusing ontologies. In *AAAI 2000 Spring Symposium on Bringing Knowledge to Business Processes*, volume 2, page 7. Karlsruhe, Germany : AAAI. <https://pdfs.semanticscholar.org/d10d/4fbbdd1d1f384f174f9611452307f4064c0ca.pdf>.
- Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries : Beyond Traditional Authority Files*. ERIC.
- Gómez-Pérez, A. (2001). Evaluation of ontologies. *International Journal of intelligent systems*, 16(3) :391–409.
- Noy, N. F., McGuinness, D. L., et al. (2001). Ontology development 101 : A guide to creating your first ontology. [http://protege.stanford.edu/publications/ontology\\_development/ontology101.pdf](http://protege.stanford.edu/publications/ontology_development/ontology101.pdf).
- Pottier, B. (2001). *Représentations mentales et catégorisations linguistiques*, volume 47. Peeters Publishers.
- Berners-Lee, T., Hendler, J., Lassila, O., et al. (2001). The semantic web. *Scientific american*, 284(5) :28–37. <https://pdfs.semanticscholar.org/566c/1c6bd366b4c9e07fc37eb372771690d5ba31.pdf>.
- Bourigault, D. et Lame, G. (2002). Analyse distributionnelle et structuration de terminologie : Application à la construction d’une ontologie documentaire du droit. *TAL. Traitement automatique des langues*, 43(1) :129–150.
- McGuinness, D. L., Fikes, R., Hendler, J., et Stein, L. A. (2002). Daml+ oil : an ontology language for the semantic web. *Intelligent Systems, IEEE*, 17(5) :72–80.
- Fernández-López, M. et Gómez-Pérez, A. (2002). The integration of ontoclean in webode. In *CEUR Workshop Proceedings*. [http://oa.upm.es/5488/1/The\\_integration\\_of.pdf](http://oa.upm.es/5488/1/The_integration_of.pdf).
- Charlet, J. (2002). L’ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales. <https://tel.archives-ouvertes.fr/tel-00006920/>.
- Rector, A. L. (2002). Normalisation of ontology implementations : Towards modularity, re-use, and maintainability. In *Proceedings Workshop on Ontologies for Multiagent Systems (OMAS) in conjunction with European Knowledge Acquisition Workshop*, pages 1–16. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.641.152&rep=rep1&type=pdf>.
- Fernández-López, M. et Gómez-Pérez, A. (2002). Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, 17(02) :129–156. <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0269888902000462>.
- Bachimont, B., Isaac, A., et Troncy, R. (2002). Semantic commitment for designing ontologies : A proposal. In *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, volume 2473 of *Lecture Notes in Computer Science*, pages 114–121. Springer Berlin Heidelberg. [http://dx.doi.org/10.1007/3-540-45810-7\\_14](http://dx.doi.org/10.1007/3-540-45810-7_14).



- 
- Laublet, P., Reynaud, C., et Charlet, J. (2002). Sur quelques aspects du web sémantique. *Actes des deuxièmes assises nationales du GdRI3*, pages 59–78. <http://www.emse.fr/~beaune/websem/03-WebSemantique.pdf>.
- Koselak, A. (2003). La sémantique naturelle d'anna wierzbicka et les enjeux interculturels. *Questions de communication*, (4) :83–95. <https://questionsdecommunication.revues.org/4611>.
- Rector, A. L. (2003). Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the 2nd international conference on Knowledge capture*, pages 121–128. ACM. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.440.4535&rep=rep1&type=pdf>.
- Rector, A., Rogers, J., Zanstra, P., et Van Der Haring, E. (2003). Opengalen : open source medical terminology and tools. In *AMIA Annual Symposium Proceedings*, volume 2003, page 982. American Medical Informatics Association. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480228/>.
- Rosse, C., Jr, J. L. M., et al. (2003). A reference ontology for biomedical informatics : the foundational model of anatomy. *Journal of biomedical informatics*, 36(6) :478–500. <https://www.ncbi.nlm.nih.gov/pubmed/14759820>.
- Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1) :99–115. [http://olst.ling.umontreal.ca/pdf/Terminology\\_2003.pdf](http://olst.ling.umontreal.ca/pdf/Terminology_2003.pdf).
- Bourigault, D., Aussenac-Gilles, N., et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18(1) :87–110. <https://pdfs.semanticscholar.org/c6f0/9760da950e577ebdb78eb1764b62c8166c03.pdf>.
- El Kalam, A. A., Deswarte, Y., Trouessin, G., et Cordonnier, E. (2004). Gestion des données médicales anonymisées : problèmes et solutions. In *2ème Conférence Francophone en Gestion et Ingenierie des Systèmes Hospitaliers (GISEH 2004)*, Mons (Belgique), pages 9–11. <https://hal.archives-ouvertes.fr/hal-00086526/>.
- Navigli, R. et Velardi, P. (2004). Learning domain ontologies from document warehouses and dedicated web sites. *Computational Linguistics*, 30(2) :151–179. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.58.124>.
- Charlet, J., Bachimont, B., et Troncy, R. (2004). Ontologies pour le web sémantique. *Revue I3*, page 31p. <https://pdfs.semanticscholar.org/f7a6/0a7db2524a88883b2303431c59475e075846.pdf>.
- Gómez-Pérez, A. (2004). Ontology evaluation. In *Handbook on ontologies*, pages 251–273. Springer. [http://link.springer.com/chapter/10.1007/978-3-540-24750-0\\_13](http://link.springer.com/chapter/10.1007/978-3-540-24750-0_13).
- McGuinness, D. L. et Van Harmelen, F. (2004). Owl web ontology, language overview. Technical report, World Wide Web Consortium. Mis à jour le 10 février 2004. [Consulté le 12.05.2017]. Disponible à l'adresse : <https://www.w3.org/TR/2004/REC-owl-features-20040210/>.

- Horridge, M., Knublauch, H., Rector, A., Stevens, R., et Wroe, C. (2004). A practical guide to building owl ontologies using the protégé-owl plugin and co-ode tools edition 1.0. *University of Manchester*. <http://people.cs.vt.edu/~kafura/ComputationalThinking/Class-Notes/Tutorial-Highlighted-Day1.pdf>.
- Smith, M. K., Welty, C., et McGuinness, D. L. (2004). Rdf 1.1 concepts and abstract syntax. Mis à jour le 10 février 2004. [Consulté le 12.05.2017]. Disponible à l'adresse : <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>.
- Bodenreider, O. (2004). The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1) :D267–D270. <https://lhncbc.nlm.nih.gov/files/archive/pub2004002.pdf>.
- Aussenac-Gilles, N. (2005). *Bottom-up methods for Knowledge Engineering*. Habilitation à diriger des recherches, Université Paul Sabatier - Toulouse III. <https://tel.archives-ouvertes.fr/tel-00089165>.
- Velardi, P., Navigli, R., Cuchiarrelli, A., et Neri, R. (2005). Evaluation of ontolearn, a methodology for automatic learning of domain ontologies. *Ontology Learning from Text : Methods, evaluation and applications*, 123 :92. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.97.333&rep=rep1&type=pdf>.
- Spaccapietra, S., Menken, M., Stuckenschmidt, H., Wache, H., Serafini, L., Tamin, A., Jarar, M., Porto, F., Parent, C., Rector, A., *et al.* (2005). Report on modularization of ontologies. *Deliverable D2*, 1(1).
- Miles, A., Matthews, B., Wilson, M., et Brickley, D. (2005). Skos core : simple knowledge organisation for the web. In *International Conference on Dublin Core and Metadata Applications*, pages pp–3.
- Cimiano, P. et Völker, J. (2005). text2onto. In *International Conference on Application of Natural Language to Information Systems*, pages 227–238. Springer.
- Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarrousse, E., Grabar, N., Ruch, P., Duff, F. L., Forget, J. F., Douyere, M., *et al.* (2005). Umlf : a unified medical lexicon for french. *International Journal of Medical Informatics*, 74(2) :119–124. <http://www.sciencedirect.com/science/article/pii/S1386505604001601>.
- Charlet, J., Bachimont, B., et Jaulent, M.-C. (2006). Building medical ontologies by terminology extraction from texts : an experiment for the intensive care units. *Computers in biology and medicine*, 36(7) :857–870. <http://www.sciencedirect.com/science/article/pii/S001048250500082X>.
- Baneyx, A. et Charlet, J. (2006). Evaluation, évolution et maintenance d'une ontologie en médecine : état des lieux et expérimentation. *Information-Interaction-Intelligence, Hors-série*. [https://www.irit.fr/journal-i3/hors\\_serie/annee2006/revue\\_i3\\_hs2006\\_01\\_07.pdf](https://www.irit.fr/journal-i3/hors_serie/annee2006/revue_i3_hs2006_01_07.pdf).
- Tsarkov, D. et Horrocks, I. (2006). Fact++ description logic reasoner : System description. In Furbach, U. et Shankar, N., editors, *Automated Reasoning*, volume 4130 of *Lecture*

- 
- Notes in Computer Science*, pages 292–297. Springer Berlin Heidelberg. [http://dx.doi.org/10.1007/11814771\\_26](http://dx.doi.org/10.1007/11814771_26).
- Aubin, S. et Hamon, T. (2006). *Improving Term Extraction with Terminological Resources*, pages 380–387. Springer Berlin Heidelberg, Berlin, Heidelberg. [http://dx.doi.org/10.1007/11816508\\_39](http://dx.doi.org/10.1007/11816508_39).
- Grau, B. C., Parsia, B., Sirin, E., et Kalyanpur, A. (2006). Modularity and web ontologies. In *Proceedings of the 10th International Conference on Principles of Knowledge Representation and Reasoning (KR2006)*, pages 198–209. <https://www.cs.ox.ac.uk/files/4567/KR06-Modularity.pdf>.
- Doran, P. (2006). Ontology reuse via ontology modularisation. In *KnowledgeWeb PhD Symposium*, volume 2006. Citeseer. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.83.9581>.
- Shadbolt, N., Hall, W., et Berners-Lee, T. (2006). The semantic web revisited. *Intelligent Systems, IEEE*, 21(3) :96–101.
- Seidenberg, J. et Rector, A. (2006). Web ontology segmentation : analysis, classification and use. In *Proceedings of the 15th international conference on World Wide Web*, pages 13–22. ACM. <http://www2006.org/programme/files/pdf/4026.pdf>.
- Baneyx, A. (2007). *Construire une ontologie de la Pneumologie Aspects théoriques, modèles et expérimentations*. PhD thesis, Université Pierre et Marie Curie-Paris VI. <https://tel.archives-ouvertes.fr/tel-00136937>.
- Uzuner, Ö., Luo, Y., et Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, 14(5) :550–563. <https://www.ncbi.nlm.nih.gov/pubmed/17600094>.
- Schvartz, S., Abi-Zeid, I., et Tourigny, N. (2007). Knowledge engineering for modelling reasoning in a diagnosis task : Application to search and rescue. *Canadian Journal of Administrative Sciences*, 24(3) :196–211. <http://dx.doi.org/10.1002/cjas.24>.
- Zeng, M., Hlava, M., Qin, J., Hodge, G., et Bedford, D. (2007). Knowledge organization systems (kos) standards. *Proceedings of the American Society for Information Science and Technology*, 44(1) :1–3. <http://onlinelibrary.wiley.com/doi/10.1002/meet.145044019/abstract>.
- Cuenca Grau, B. et Kutz, O. (2007). Modular ontology languages revisited. In *Proc. of the Workshop on Semantic Web for Collaborative Knowledge Acquisition*. <https://www.inf.unibz.it/~okutz/resources/MOL.pdf>.
- Peace, J. et Brennan, P. (2007). Ontological representation of family and family history. In *AMIA Annual Symposium proceedings*, page 1072. <https://www.ncbi.nlm.nih.gov/pubmed/18694170>.
- Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., et Katz, Y. (2007). Pellet : A practical owl-dl reasoner. *Web Semantics : science, services and agents on the World Wide Web*, 5(2) :51–53. <http://www.sciencedirect.com/science/article/pii/S1570826807000169>.

- Stenzhorn, H., Beibwanger, E., Schulz, S., *et al.* (2007). Towards a top-domain ontology for linking biomedical ontologies. In *Medinfo 2007 : Proceedings of the 12th World Congress on Health (Medical) Informatics ; Building Sustainable Health Systems*, page 1225. IOS Press. <https://www.ncbi.nlm.nih.gov/pubmed/17911910>.
- Völker, J., Vrandečić, D., Sure, Y., et Hotho, A. (2008). Aeon—an approach to the automatic evaluation of ontologies. *Applied Ontology*, 3(1) :41–62.
- Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., et Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC medical informatics and decision making*, 8(1) :1. <https://dspace.mit.edu/handle/1721.1/41622>.
- MeSH (2008). Définition du terme « glossaire » par le medical subject headings. <http://mesh.inserm.fr/mesh/view/loadSheet.jsp?sheetId=D016437>, visité le 04-02-2016.
- Shearer, R., Motik, B., et Horrocks, I. (2008). Hermit : A highly-efficient owl reasoner. In *OWLED*, volume 432. <https://www.cs.ox.ac.uk/boris.motik/pubs/smh08Hermit.pdf>.
- Demazeux, S. (2008). Les catégories psychiatriques sont-elles dépassées? *Psychiatrie Sciences Humaines Neurosciences*, 6(1) :17–25. <https://doi.org/10.1007/2Fs11836-008-0049-z>.
- D'Aquin, M., Haase, P., Rudolph, S., Euzenat, J., Zimmermann, A., Dzbor, M., Iglesias, M., Jacques, Y., Caracciolo, C., Buil Aranda, C., et Jose Manuel, G. (2008). NeOn Formalisms for Modularization : Syntax, Semantics, Algebra. Research report, Universität Karlsruhe. <https://hal.archives-ouvertes.fr/hal-01242833>.
- Haase, P., Lewen, H., Studer, R., Tran, D. T., Erdmann, M., d'Aquin, M., et Motta, E. (2008). The neon ontology engineering toolkit. [http://www.aifb.kit.edu/images/7/7e/2008\\_1757\\_Haase\\_The\\_NeOn\\_Ontolo\\_1.pdf](http://www.aifb.kit.edu/images/7/7e/2008_1757_Haase_The_NeOn_Ontolo_1.pdf).
- Group, W. O. W. (2008). Owl 2 web ontology language : Profiles. Technical report, World Wide Web Consortium. Mis à jour le 08 octobre 2008. [Consulté le 12.05.2017]. Disponible à l'adresse : <https://www.w3.org/TR/2008/WD-owl2-profiles-20081008/>.
- Prud'Hommeaux, E., Seaborne, A., *et al.* (2008). Sparql query language for rdf. *W3C recommendation*, 15. <https://www.w3.org/TR/rdf-sparql-query/>.
- Möller, H.-J. (2008). Systematic of psychiatric disorders between categorical and dimensional approaches. *European Archives of Psychiatry and Clinical Neuroscience*, 258(2) :48–73. <https://link.springer.com/article/10.1007/s00406-008-2004-3>.
- Aussenac-Gilles, N., Despres, S., et Szulman, S. (2008). The terminae method and platform for ontology engineering from texts. *Bridging the Gap between Text and Knowledge-Selected Contributions to Ontology Learning and Population from Text*, pages 199–223. <https://hal.archives-ouvertes.fr/hal-00174388>.

- 
- Beckett, D. et Berners-Lee, T. (2008). Turtle-terse rdf triple language. Mis à jour le 14 juillet 2008. [Consulté le 12.05.2017]. Disponible à l'adresse : <https://www.w3.org/TeamSubmission/2008/SUBM-turtle-20080114/>.
- Mazuel, L. et Charlet, J. (2009). Aligement entre des ontologies de domaine et la snomed : trois études de cas. In *20ème conférence sur l'Ingénierie des Connaissances-IC2009*, pages A–paraître. <https://hal.archives-ouvertes.fr/file/index/docid/377516/filename/spim09-revisedV2.pdf>.
- Szulman, S., Charlet, J., Aussenac-Gilles, N., Nazarenko, A., Sardet, É., et Téguiak, H. V. (2009). Dafoe : an ontology building platform from text or thesauri. In *International Conference on Knowledge Engineering and Ontology Development (KEOD 2009)*, pages 1–4. <https://hal.archives-ouvertes.fr/hal-00525527/en/>.
- Hitzler, P., Krotzsch, M., et Rudolph, S. (2009). Knowledge representation for the semantic web. <http://www.semantic-web-book.org/w/images/b/b0/KI09-OWL-Rules-1.pdf>.
- Reymonet, A., Thomas, J., Aussenac-Gilles, N., et al. (2009). Modélisation de ressources termino-ontologiques en owl. *Actes des Journées Francophones d'Ingénierie des Connaissances (IC 2007)*, pages 169–180. <https://hal.archives-ouvertes.fr/hal-00365888>.
- Lenne, M. D. (2009). Modélisation des connaissances et de l'interaction. *Application aux Environnements Informatiques pour l'Apprentissage Humain*.
- Vrandečić, D. (2009). *Ontology evaluation*. Springer. [http://link.springer.com/chapter/10.1007/978-3-540-92673-3\\_13](http://link.springer.com/chapter/10.1007/978-3-540-92673-3_13).
- Guarino, N. et Welty, C. (2009). An overview of ontoclean. In Staab, S. et Studer, R., editors, *Handbook on Ontologies*, International Handbooks on Information Systems, pages 201–220. Springer Berlin Heidelberg. [http://dx.doi.org/10.1007/978-3-540-92673-3\\_9](http://dx.doi.org/10.1007/978-3-540-92673-3_9).
- Ji, Q., Haase, P., Qi, G., Hitzler, P., et Stadtmüller, S. (2009). Radon repair and diagnosis in ontology networks. In *The semantic web : research and applications*, pages 863–867. Springer. [https://link.springer.com/chapter/10.1007/978-3-642-02121-3\\_71](https://link.springer.com/chapter/10.1007/978-3-642-02121-3_71).
- Pérez, J., Arenas, M., et Gutierrez, C. (2009). Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, 34(3) :16. <http://iswc2006.semanticweb.org/items/Arenas2006bv.pdf>.
- Pathak, J., Johnson, T. M., et Chute, C. G. (2009). Survey of modular ontology techniques and their applications in the biomedical domain. *Integrated computer-aided engineering*, 16(3) :225. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3113511/>.
- Grouin, C., Rosier, A., Dameron, O., et Zweigenbaum, P. (2009). Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. *Risques, technologies de l'information pour les pratiques médicales*, pages 23–34.



- Pies, R. (2009). What should count as a mental disorder in dsm-v? *Psychiatric Times*, 26(4) :17–17. <http://www.psychiatrictimes.com/dsm-5-0/what-should-count-mental-disorder-dsm-v>.
- Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., et Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record : a review of recent research. *BMC medical research methodology*, 10(1) :1. <https://www.ncbi.nlm.nih.gov/pubmed/20678228>.
- Pammer, V., Ghidini, C., Rospocher, M., Serafini, L., et Lindstaedt, S. (2010). Automatic support for formative ontology evaluation. In *Poster Proceedings of the Conference on Knowledge Engineering and Knowledge Management by the Masses (EKAW-10)*. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.204.5558>.
- Dhombres, F., Jouannic, J.-M., Jaulent, M.-C., et Charlet, J. (2010). Choix méthodologiques pour la construction d’une ontologie de domaine en médecine périnatale. In *21èmes Journées Francophones d’Ingénierie des Connaissances*, pages 171–182. Ecole des Mines d’Alès. <https://hal.archives-ouvertes.fr/hal-00487736/>.
- Szulman, S., Charlet, J., Aussenac-Gilles, N., Nazarenko, A., Hernandez, N., Nada, N., Sardet, E., Delahousse, J., et Téguia, H. (2010). Dafoe : une plateforme multi-méthodes et multi-modèles pour construire des ontologies de domaine (demo). *Congrès Francophone de Reconnaissance des Formes et Intelligence Artificielle, RFIA*, 10 :1–2.
- Denis, P. et Sagot, B. (2010). Exploitation d’une ressource lexicale pour la construction d’un étiqueteur morphosyntaxique état-de-l’art du français. In *Traitement Automatique des Langues Naturelles : TALN 2010*. <https://hal.inria.fr/inria-00521231/>.
- Ceusters, W. et Smith, B. (2010). Foundations for a realist ontology of mental disease. *Journal of Biomedical Semantics*, 1(1) :10. <http://dx.doi.org/10.1186/2041-1480-1-10>.
- Aussenac-Gilles, N. et Charlet, J. (2010). Ingénierie des connaissances modélisation (2).
- Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., et Wang, P. (2010). Research domain criteria (rdoc) : toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, 167(7) :748–751. <http://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2010.09091379>.
- Cartoni, B. et Zweigenbaum, P. (2010). Semi-automated extension of a specialized medical lexicon for french. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*. Valletta, Malta. [https://perso.limsi.fr/pz/FTPapiers/Cartoni\\_LREC2010.pdf](https://perso.limsi.fr/pz/FTPapiers/Cartoni_LREC2010.pdf).
- Kola, J. S., Harris, J., Lawrie, S., Rector, A., Goble, C., et Martone, M. (2010). Towards an ontology for psychosis. *Cognitive Systems Research*, 11(1) :42–52. <http://www.sciencedirect.com/science/article/pii/S1389041708000375>.
- Özacar, T., Öztürk, Ö., et Ünalır, M. O. (2011). Anemone : An environment for modular ontology development. *Data & Knowledge Engineering*, 70(6) :504–526. <https://doi.org/10.1016/j.datak.2011.02.005>.

- 
- Vandenbussche, P.-Y. (2011). *Definition of a formal framework for Knowledge Organization Systems representation*. Theses, Université Pierre et Marie Curie - Paris VI. <https://tel.archives-ouvertes.fr/tel-00642545>.
- Everaert-Desmedt, N. (2011). La sémiotique de peirce [en ligne]. Mis à jour en 2011. [Consulté le 12.05.2017]. Disponible à l'adresse : <http://www.signosemio.com/peirce/semiotique.asp>.
- Aimé, X. (2011). *Prototypicality gradients, similarity and proximity measures : a contribution to the Ontology Engineering*. Theses, Université de Nantes. <https://tel.archives-ouvertes.fr/tel-00660916>.
- Rocheteau, J. et Daille, B. (2011). Ttc termsuite : A uima application for multilingual terminology extraction from comparable corpora. In *5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 9–12. [https://hal.archives-ouvertes.fr/file/index/docid/819025/filename/TTC\\_IJCNLP\\_2011.pdf](https://hal.archives-ouvertes.fr/file/index/docid/819025/filename/TTC_IJCNLP_2011.pdf).
- Hamon, T. (2012). Acquisition terminologique pour identifier les mots clés d'articles scientifiques. *Actes du huitième DÉfi Fouille de Textes*, page 28. <http://www.aclweb.org/anthology/W12-1103>.
- Misès, R. (2012). Classification française des troubles mentaux de l'enfant et de l'adolescent–r-2012. *Correspondance et transcodage CIM10. 5e édition*. Rennes : Presse de l'EHESP.
- Omrane, N., Nazarenko, A., et Szulman, S. (2012). Comment guider le travail de normalisation terminologique? In *23es Journées francophones d'Ingénierie des Connaissances. (IC 2012)*, page poster, Paris, France. <https://hal.archives-ouvertes.fr/hal-00704294>.
- Charlet, J., Declerck, G., Dhombres, F., Gayet, P., Miroux, P., et Vandenbussche, P.-Y. (2012). Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation. In *IC-23èmes Journées francophones d'Ingénierie des Connaissances*, pages 33–48. <http://www.hal.inserm.fr/hal-00717807/>.
- Denis, P. et Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language resources and evaluation*, 46(4) :721–736. <http://link.springer.com/article/10.1007/s10579-012-9193-0>.
- Hébert, L. et Dumont-Morin, G. (2012). Dictionnaire de sémiotique générale. *Louis Hébert, (éd.). Signo, éd. Louis Hébert, Rimouski, Québec*. <http://www.signosemio.com/documents/dictionnaire-semiotique-generale.pdf>.
- Lange, C., Kutz, O., Mossakowski, T., et Grüninger, M. (2012). The distributed ontology language (dol) : ontology integration and interoperability applied to mathematical formalization. In *International Conference on Intelligent Computer Mathematics*, pages 463–467. Springer.

- The Dublin Core Metadata Initiative - DCMI (2012). Dublin core metadata element set, version 1.1. Mis à jour le 14 juin 2012. [Consulté le 08.04.2016]. Disponible à l'adresse : <http://dublincore.org/documents/dces/>.
- Dublin Core Qualifiers (2012). Dublin core metadata element set, version 1.1. Mis à jour le 11 juillet 2000. [Consulté le 08.04.2016]. Disponible à l'adresse : <http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/>.
- Gicquel, Q., Proux, D., Marchal, P., Hagège, C., Berrouane, Y., Darmoni, S. J., Pereira, S., Segond, F., et Metzger, M.-H. (2012). Évaluation d'un outil d'aide à l'anonymisation des documents médicaux basé sur le traitement automatique du langage naturel. *Systèmes d'information pour l'amélioration de la qualité en santé*, pages 165–176. [http://link.springer.com/chapter/10.1007/978-2-8178-0285-5\\_15](http://link.springer.com/chapter/10.1007/978-2-8178-0285-5_15).
- Alecu, B. P., Thomas, I., et Renahy, J. (2012). La "multi-extraction" comme stratégie d'acquisition optimisée de ressources terminologiques et non terminologiques. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, pages 511–518, Grenoble, France. [http://www.aclweb.org/website/old\\_anthology/F/F12/F12-2047.pdf](http://www.aclweb.org/website/old_anthology/F/F12/F12-2047.pdf).
- Ghidini, C., Rospocher, M., et Serafini, L. (2012). Modeling in a wiki with moki : Reference architecture, implementation, and usages. *International Journal On Advances in Life Sciences*, 4(3 et 4) :111–124. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.674.1541&rep=rep1&type=pdf>.
- Schober, D., Tudose, I., Svatek, V., et Boeker, M. (2012). Ontocheck : verifying ontology naming conventions and metadata completeness in protege 4. *Journal of Biomedical Semantics*, 3(Suppl 2) :S4. <http://www.jbiomedsem.com/content/3/S2/S4>.
- Sabou, M. et Fernandez, M. (2012). Ontology (network) evaluation. In *Ontology Engineering in a Networked World*, pages 193–212. Springer Berlin Heidelberg. [http://dx.doi.org/10.1007/978-3-642-24794-1\\_9](http://dx.doi.org/10.1007/978-3-642-24794-1_9).
- Aimé, X. et Charlet, J. (2012). Preferred label validation by lexical prototypicality gradient : a use case on a rare diseases ontology. *on Capturing and Refining Knowledge in the Medical Domain (K-MED 2012)*, page 36.
- Gherasim, T., Berio, G., Harzallah, M., Kuntz, P., *et al.* (2012). Problems impacting the quality of automatically built ontologies. *Knowledge Engineering and Software Engineering (KESE8)*, page 22.
- Declerck, G., Baneyx, A., Aimé, X., et Charlet, J. (2012). A quoi servent les ontologies fondationnelles ? In *23èmes Journées francophones d'Ingénierie des Connaissances (IC 2012)*, pages pp–67. <https://hal.archives-ouvertes.fr/hal-00714656/>.
- Hastings, J., Smith, B., Ceusters, W., Jensen, M., et Mulligan, K. (2012). Representing mental functioning : Ontologies for mental health and disease. In *ICBO 2012 : 3rd International Conference on Biomedical Ontology*. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.308.236&rep=rep1&type=pdf>.



- 
- Morris, S. E. et Cuthbert, B. N. (2012). Research domain criteria : cognitive systems, neural circuits, and dimensions of behavior. *Dialogues Clin Neurosci*, 14(1) :29–37. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.270.1468&rep=rep1&type=pdf>.
- Misès, R., Bursztejn, C., Botbol, M., Coincon, Y., Durand, B., Garrabe, J., Garret-Gloanec, N., Golse, B., Portelli, C., Raynaud, J.-P., *et al.* (2012). Une nouvelle version de la classification française des troubles mentaux de l'enfant et de l'adolescent : la cftmea r 2012, correspondances et transcodages avec l'icd 10. *Neuropsychiatrie de l'enfance et de l'adolescence*, 60(6) :414–418. [https://projet.chu-besancon.fr/pmb/PMB\\_Ecoles/opac\\_css/doc\\_num.php?explnum\\_id=261](https://projet.chu-besancon.fr/pmb/PMB_Ecoles/opac_css/doc_num.php?explnum_id=261).
- Poveda-Villalón, M., Suárez-Figueroa, M. C., et Gómez-Pérez, A. (2012). Validating ontologies with OOPS! In *Knowledge Engineering and Knowledge Management*, pages 267–281. Springer. [http://link.springer.com/chapter/10.1007/978-3-642-33876-2\\_24](http://link.springer.com/chapter/10.1007/978-3-642-33876-2_24).
- Richard, M., Aimé, X., Krebs, M.-O., et Charlet, J. (2013). Au delà du dsm : les ontologies comme aide aux classifications descriptives psychiatriques? In *2e édition du Symposium sur l'Ingénierie de l'Information Médicale*. <http://hal.upmc.fr/hal-00840700/>.
- Grouin, C. (2013). *Clinical Records De-Identification : Performances and Limits of Rule-based and Machine-Learning based Approaches*. Theses, Université Pierre et Marie Curie - Paris VI. <https://tel.archives-ouvertes.fr/tel-00848672>.
- Lossio-Ventura, J. A., Jonquet, C., Roche, M., et Teisseire, M. (2013). Combining c-value and keyword extraction methods for biomedical terms extraction. In *LBM : Languages in Biology and Medicine*. <https://hal.archives-ouvertes.fr/lirmm-01019991/>.
- Ressad-Boudighaghen, O., Szulman, S., Zargayouna, H., et Paul, E. (2013). Construction collaborative d'une ressource termino-ontologique (rto) pour le droit des collectivités territoriales. In *IC-24èmes Journées francophones d'Ingénierie des Connaissances*, IC 2013, Lille, France. <https://hal.archives-ouvertes.fr/hal-00860104>.
- APA *et al.* (2013). *Diagnostic and statistical manual of mental disorders (DSM-5)*.
- Aimé, X. et Charlet, J. (2013). Ic : Ingénierie des connaissances ou ingénierie du conformisme? In *IC-24èmes Journées francophones d'Ingénierie des Connaissances*. <https://hal.archives-ouvertes.fr/hal-01103767>.
- Stern, R. (2013). *Identification automatique d'entités pour l'enrichissement de contenus textuels*. PhD thesis, Université Paris-Diderot-Paris VII. <https://hal.archives-ouvertes.fr/tel-00939420/>.
- Desfriches Doria, O. (2013). *La classification à facettes pour la gestion des connaissances métier : méthodologie d'élaboration de FolkClassifications à facettes*. PhD thesis. <http://www.theses.fr/2013CNAM0903/document>.

- Widakowich, C., Van Wettere, L., Jurysta, F., Linkowski, P., et Hubain, P. (2013). L'approche dimensionnelle versus l'approche catégorielle dans le diagnostic psychiatrique : aspects historiques et épistémologiques. In *Annales Médico-psychologiques, revue psychiatrique*, volume 171, pages 300–305. Elsevier. <http://www.sciencedirect.com/science/article/pii/S0003448712002259>.
- Velardi, P., Faralli, S., et Navigli, R. (2013). Ontolearn reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3) :665–707. [http://www.mitpressjournals.org/doi/abs/10.1162/COLI\\_a\\_00146](http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00146).
- Frances, A. *et al.* (2013). Saving normal : An insider's revolt against out-of-control psychiatric diagnosis, dsm-5, big pharma and the medicalization of ordinary life. *Psychotherapy in Australia*, 19(3) :14.
- Quantin, C., Allaert, F., Auverlot, B., et Rialle, V. (2013). Sécurité, aspects juridiques et éthiques des données de santé informatisées. In *Informatique médicale, e-Santé*, pages 265–305. Springer.
- Harris, S., Seaborne, A., et Prud'hommeaux, E. (2013). Sparql 1.1 query language. *W3C Recommendation*, 21. <https://www.w3.org/TR/sparql11-query/>.
- Mattatia, F. (2013). *Traitement des données personnelles : Le guide juridique-La loi Informatique et libertés et la CNIL-Jurisprudences*. Editions Eyrolles.
- Ben Abacha, A., Da Silveira, M., et Pruski, C. (2013). Une approche pour la validation du contenu d'une ontologie par un système à base de questions/réponses. In *IC - 24èmes Journées francophones d'Ingénierie des Connaissances*, Lille, France. <https://hal.inria.fr/hal-01103777>.
- Tudorache, T., Nyulas, C., Noy, N. F., et Musen, M. A. (2013). Webprotégé : A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web*, 4(1) :89–99. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691821/>.
- Del Vescovo, C., Hahmann, T., Pearce, D., et Walther, D. (2013). Workshop on modular ontologies (womo) 2013. [http://www.ceur-ws.org/Vol-1081/womo2013\\_proceedings.pdf](http://www.ceur-ws.org/Vol-1081/womo2013_proceedings.pdf).
- Lossio-Ventura, J. A., Jonquet, C., Roche, M., et Teisseire, M. (2014). Biotex : A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the ISWC 2014 - the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014*, pages 157–160. <https://hal.archives-ouvertes.fr/hal-01136531>.
- Drame, K. (2014). *Contribution à la construction d'ontologies et à la recherche d'information : application au domaine médical*. PhD thesis, Université de Bordeaux. <https://tel.archives-ouvertes.fr/tel-01166042/>.
- Charlet, J., Mazuel, L., Declerck, G., Miroux, P., et Gayet, P. (2014). Describing localized diseases in medical ontology : an fma-based algorithm. In *MIE*, pages 1023–1027. <https://www.ncbi.nlm.nih.gov/pubmed/25160343>.

- 
- Silva, C., Marreiros, G., et Silva, N. (2014). Development of an ontology for supporting diagnosis in psychiatry. In *Distributed Computing and Artificial Intelligence, 11th International Conference*, pages 343–350. Springer. <https://www.ncbi.nlm.nih.gov/pubmed/19697514>.
- Picard, R. (2014). Médecine personnalisée : de quoi parle-t-on ? une vision prospective. In *Annales des Mines-Réalités industrielles*, number 4, pages 99–106. ESKA. <https://www.cairn.info/revue-realites-industrielles1-2014-4-page-99.htm>.
- Insel, T. R. (2014). The nimh research domain criteria (rdoc) project : precision medicine for psychiatry. *American Journal of Psychiatry*, 171(4) :395–397. <http://ajp.psychiatryonline.org/doi/full/10.1176/appi.ajp.2014.14020138>.
- Bourgin, J. et Duchesnay, E. (2014). Phénotypes cliniques précoces et recherche de biomarqueurs stratégiques : les fondements d’une psychiatrie personnalisée. *L’information psychiatrique*, 89(10) :781–789. <http://www.cairn.info/revue-l-information-psychiatrique-2013-10-page-781.html>.
- Cuthbert, B. N. (2014). The rdoc framework : facilitating transition from icd/dsm to dimensional approaches that integrate neuroscience and psychopathology. *World Psychiatry*, 13(1) :28–35. <http://onlinelibrary.wiley.com/doi/10.1002/wps.20087/full>.
- Dramé, K., Diallo, G., Delva, F., Dartigues, J. F., Mouillet, E., Salamon, R., et Mougin, F. (2014). Reuse of termino-ontological resources and text corpora for building a multilingual domain ontology : an application to alzheimer’s disease. *Journal of biomedical informatics*, 48 :171–182. <http://www.sciencedirect.com/science/article/pii/S1532046413002049>.
- Lossio-Ventura, J. A., Jonquet, C., Roche, M., et Teisseire, M. (2014). Towards a mixed approach to extract biomedical terms from text corpus. *International Journal of Knowledge Discovery in Bioinformatics*, 4(1) :1–15. <https://hal-lirmm.ccsd.cnrs.fr/lirmm-00859846/>.
- Héon, M. (2014). *Web sémantique et modélisation ontologique (avec G-OWL)*. Editions ENI.
- Phillips, M. R. (2014). Will rdoc hasten the decline of america’s global leadership role in mental health? *World Psychiatry*. <http://onlinelibrary.wiley.com/doi/10.1002/wps.20098/full>.
- Lossio-Ventura, J. A., Jonquet, C., Roche, M., et Teisseire, M. (2014). Yet another ranking function for automatic multiword term extraction. In *International Conference on Natural Language Processing*, pages 52–64. Springer. <https://hal.archives-ouvertes.fr/lirmm-01068556>.
- BnF (2015). Bnf - dublin core [en ligne]. Mis à jour le 21 décembre 2016. [Consulté le 12.05.2017]. Disponible à l’adresse : [http://www.bnf.fr/fr/professionnels/formats\\_catalogue/a.f\\_dublin\\_core.html](http://www.bnf.fr/fr/professionnels/formats_catalogue/a.f_dublin_core.html).

- Aimé, X. (2015). Eléments de réflexion sur l'utilisation de corpus pour la construction d'ontologies. In *IC2015*. <https://hal.archives-ouvertes.fr/hal-01167550/>.
- Musen, M. A. (2015). The protégé project : A look back and a look forward. *AI matters*, 1(4) :4–12. <http://dl.acm.org/citation.cfm?id=2757003>.
- Yee, C. M., Javitt, D. C., et Miller, G. A. (2015). Replacing dsm categorical analyses with dimensional analyses in psychiatry research : The research domain criteria initiative. *JAMA psychiatry*, 72(12) :1159–1160. <https://www.ncbi.nlm.nih.gov/pubmed/26559005>.
- Cuthbert, B. N. (2015). Research domain criteria : toward future psychiatric nosologies. *Dialogues Clin Neurosci*, 17(1) :89–97. <http://pubmedcentralcanada.ca/pmcc/articles/PMC4421905/>.
- Weinberger, D. R., Glick, I. D., et Klein, D. F. (2015). Whither research domain criteria (rdoc) ? : The good, the bad, and the ugly. *JAMA psychiatry*, 72(12) :1161–1162. <http://jamanetwork.com/journals/jamapsychiatry/article-abstract/2469109>.
- Lilienfeld, S. O. et Treadway, M. T. (2016). Clashing diagnostic approaches : Dsm-icd versus rdoc. *Annual review of clinical psychology*, 12 :435–463. <https://www.ncbi.nlm.nih.gov/pubmed/26845519>.
- Young, G. (2016). The dsm-5 and the rdoc : Grand designs and grander problems. In *Unifying Causality and Psychology*, pages 591–610. Springer. [http://link.springer.com/chapter/10.1007/978-3-319-24094-7\\_23](http://link.springer.com/chapter/10.1007/978-3-319-24094-7_23).
- Maaroufi, M. (2016). *Interopérabilité des données médicales dans le domaine des maladies rares dans un objectif de santé publique*. PhD thesis, Université Pierre et Marie Curie - Paris 6. <https://tel.archives-ouvertes.fr/tel-01446534>.
- Steinberg, K. (2016). Qualité des données de santé disponibles en France et de leurs modèles - Comment la garantir pour répondre aux enjeux de la gestion des connaissances médicales ? Master's thesis, INTD-CNAM-Institut national des techniques de la documentation. [https://memsic.ccsd.cnrs.fr/mem\\_01476178](https://memsic.ccsd.cnrs.fr/mem_01476178).

---

## **Troisième partie**

### **Annexes**



## Cadre administratif de la thèse

La thèse qui fait l'objet de ce manuscrit a été réalisée au sein du Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé<sup>1</sup> (LIMICS), URM\_S 1142, de l'Institut National de la Santé Et de la Recherche Médicale<sup>2</sup> (INSERM). Ce travail s'est déroulé de septembre 2013 à juin 2017 et a été financé par une bourse doctorale attribuée par l'école doctorale 393 Pierre Louis de santé publique : épidémiologie et sciences de l'information biomédicale<sup>3</sup> de l'Université Pierre et Marie Curie<sup>4</sup> (Paris 6).

---

1. <http://www.limics.fr/>  
2. <http://www.inserm.fr/>  
3. <http://www.ed393.upmc.fr/>  
4. <http://www.upmc.fr/>





# Illustration des théories de la sémiotique de Hébert et Dumont-Morin [2012]

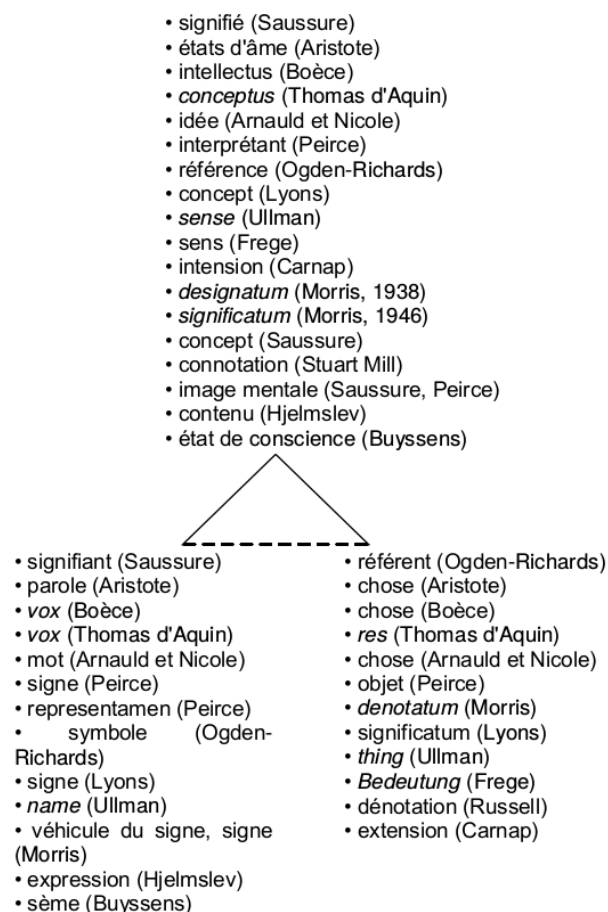


FIGURE B.1 – Les diverses appellations du signe linguistique, page 248 de Hébert et Dumont-Morin [2012].



# Illustration de la dixième version de la Classification statistique Internationale des Maladies et des problèmes de santé connexes

CIM-10 Version 2008

**Chapitre V**  
**Troubles mentaux et du comportement**  
**(F00-F99)**

**Inclus:** troubles du développement psychologique

**Excl.:** symptômes, signes et résultats anormaux d'examens cliniques et de laboratoire, non classés ailleurs (R00-R99)

**Ce chapitre comprend les blocs suivants :**

- F00-F09 Troubles mentaux organiques, y compris les troubles symptomatiques
- F10-F19 Troubles mentaux et du comportement liés à l'utilisation de substances psycho-actives
- F20-F29 Schizophrénie, trouble schizotypique et troubles délirants
- F30-F39 Troubles de l'humeur [affectifs]
- F40-F48 Troubles névrotiques, troubles liés à des facteurs de stress et troubles somatoformes
- F50-F59 Syndromes comportementaux associés à des perturbations physiologiques et à des facteurs physiques
- F60-F69 Troubles de la personnalité et du comportement chez l'adulte
- F70-F79 Retard mental
- F80-F89 Troubles du développement psychologique
- F90-F98 Troubles du comportement et troubles émotionnels apparaissant habituellement durant l'enfance et l'adolescence
- F99-F99 Trouble mental, sans précision

**Les catégories de ce chapitre comprenant des astérisques sont les suivantes :**

- F00\* Démence de la maladie d'Alzheimer
- F02\* Démence au cours d'autres maladies classées ailleurs

FIGURE C.1 – Cette image illustre la description d'une catégorie dans la CIM-10. Un code est associé à un libellé et à des indications thérapeutiques. Par exemple sur cette image, le code F00\* est associé au libellé « Démence de la maladie d'Alzheimer ». Le code entre parenthèse G30.0-+ indique un lien avec le code diagnostique G30.0 qui est celui associé au libellé « Maladie d'Alzheimer à début précoce »

## **CHAPITRE C : Illustration de la dixième version de la Classification statistique Internationale des Maladies et des problèmes de santé connexes**

La démence (F00-F03) est un syndrome dû à une maladie cérébrale, habituellement chronique et progressive, caractérisé par une altération de nombreuses fonctions corticales supérieures, telles que la mémoire, l'idéation, l'orientation, la compréhension, le calcul, la capacité d'apprendre, le langage et le jugement. Le syndrome ne s'accompagne pas d'un obscurcissement de la conscience. Les déficiences des fonctions cognitives s'accompagnent habituellement (et sont parfois précédées) d'une détérioration du contrôle émotionnel, du comportement social, ou de la motivation. Ce syndrome survient dans la maladie d'Alzheimer, dans les maladies vasculaires cérébrales, et dans d'autres affections qui de manière primaire ou secondaire, affectent le cerveau.

Utiliser, au besoin, un code supplémentaire, pour identifier la maladie sous-jacente.

<b>F00*</b>	<b>Démence de la maladie d'Alzheimer (G30.-+)</b>
	La maladie d'Alzheimer est une maladie cérébrale dégénérative primitive d'étiologie inconnue dont la neuropathologie et la neurochimie sont caractéristiques. Elle débute habituellement de façon insidieuse et progresse lentement mais régulièrement en quelques années.
<b>F00.0*</b>	<b>Démence de la maladie d'Alzheimer, à début précoce (G30.0+)</b>
	Démence de la maladie d'Alzheimer survenant avant l'âge de 65 ans, évoluant assez rapidement vers une détérioration et comportant de multiples perturbations marquées des fonctions corticales supérieures.
	Démence dégénérative primaire de type Alzheimer, à début présénile Démence présénile, de type Alzheimer Maladie d'Alzheimer, type 2
<b>F00.1*</b>	<b>Démence de la maladie d'Alzheimer, à début tardif (G30.1+)</b>
	Démence de la maladie d'Alzheimer survenant après l'âge de 65 ans, habituellement à la fin de la huitième décennie ou au-delà ; elle évolue de façon lentement progressive et se caractérise essentiellement par une altération de la mémoire.
	Démence dégénérative primaire de type Alzheimer, à début sénile Démence sénile, de type Alzheimer (DSTA) Maladie d'Alzheimer, type 1
<b>F00.2*</b>	<b>Démence de la maladie d'Alzheimer, forme atypique ou mixte (G30.8+)</b>
	Démence atypique, de type Alzheimer
<b>F00.9*</b>	<b>Démence de la maladie d'Alzheimer, sans précision (G30.9+)</b>

FIGURE C.2 – Cette image illustre les différentes catégories regroupés dans le Chapitre 5 de la CIM-10 : « Troubles mentaux et du comportement ». Chacune de ces catégories principales est désigné par un intervalle de codes.

## Les « Constructs/Subconstructs » des RDoC

### 1. Domain : Negative Valence Systems

- Construct : Acute Threat ("Fear")
- Construct : Potential Threat ("Anxiety")
- Construct : Sustained Threat
- Construct : Loss
- Construct : Frustrative Nonreward

### 2. Domain : Positive Valence Systems

- Construct : Approach Motivation
  - Subconstruct : Reward Valuation
  - Subconstruct : Effort Valuation / Willingness to Work
  - Subconstruct : Expectancy / Reward Prediction Error
  - Subconstruct : Action Selection / Preference-Based Decision Making
- Construct : Initial Responsiveness to Reward Attainment
- Construct : Sustained/Longer-Term Responsiveness to Reward Attainment
- Construct : Reward Learning
- Construct : Habit

### 3. Domain : Cognitive Systems

- Construct : Attention
- Construct : Perception
  - Subconstruct : Visual Perception
  - Subconstruct : Auditory Perception
  - Subconstruct : Olfactory/Somatosensory/Multimodal/Perception
- Construct : Declarative Memory
- Construct : Language

- Construct : Cognitive Control
    - Subconstruct : Goal Selection ; Updating, Representation, and Maintenance
    - Subconstruct : Response Selection ; Inhibition/Suppression
    - Subconstruct : Performance Monitoring
  - Construct : Working Memory
    - Subconstruct : Active Maintenance
    - Subconstruct : Flexible Updating
    - Subconstruct : Limited Capacity
    - Subconstruct : Interference Control
4. Domain : Social Processes
- Construct : Affiliation and Attachment
  - Construct : Social Communication
    - Subconstruct : Reception of Facial Communication
    - Subconstruct : Production of Facial Communication
    - Subconstruct : Reception of Non-Facial Communication
    - Subconstruct : Production of Non-Facial Communication
  - Construct : Perception and Understanding of Self
    - Subconstruct : Agency
    - Subconstruct : Self-Knowledge
  - Construct : Perception and Understanding of Others
    - Subconstruct : Animacy Perception
    - Subconstruct : Action Perception
    - Subconstruct : Understanding Mental States
5. Domain : Arousal and Regulatory Systems
- Construct : Arousal
  - Construct : Circadian Rhythms
  - Construct : Sleep-Wakefulness

## Nuages des liens entre les données ouvertes

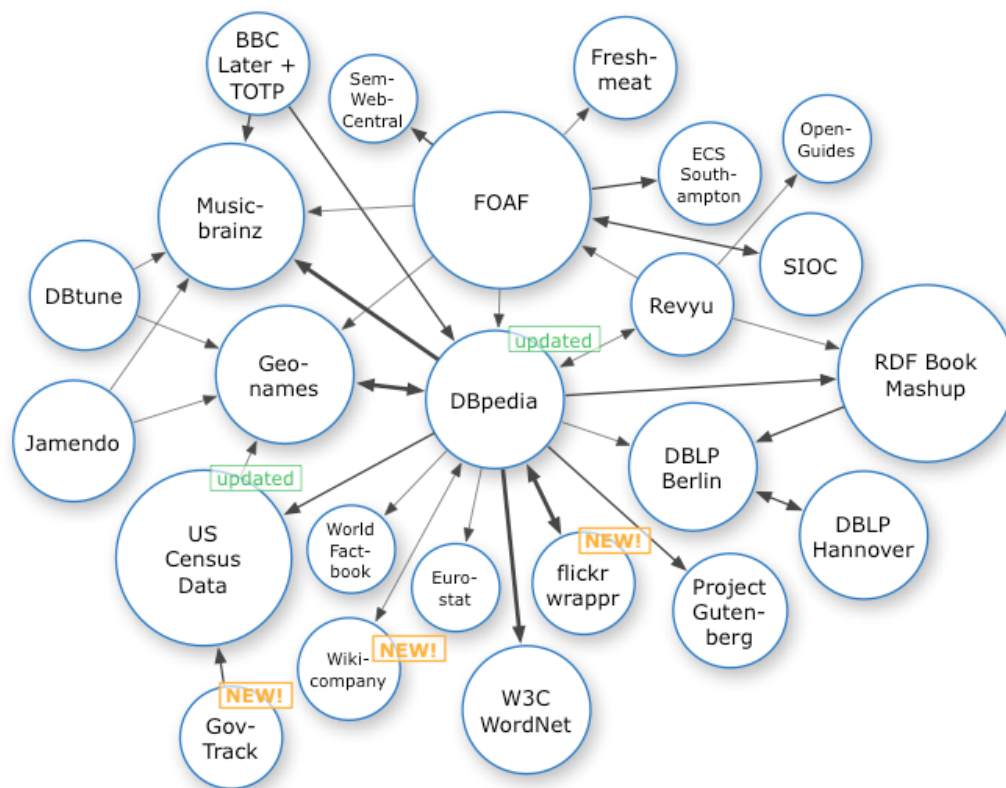


FIGURE E.1 – Les données liées en 2007.





FIGURE E.2 – Les données liées en 2009.



FIGURE E.3 – Les données liées en 2010.

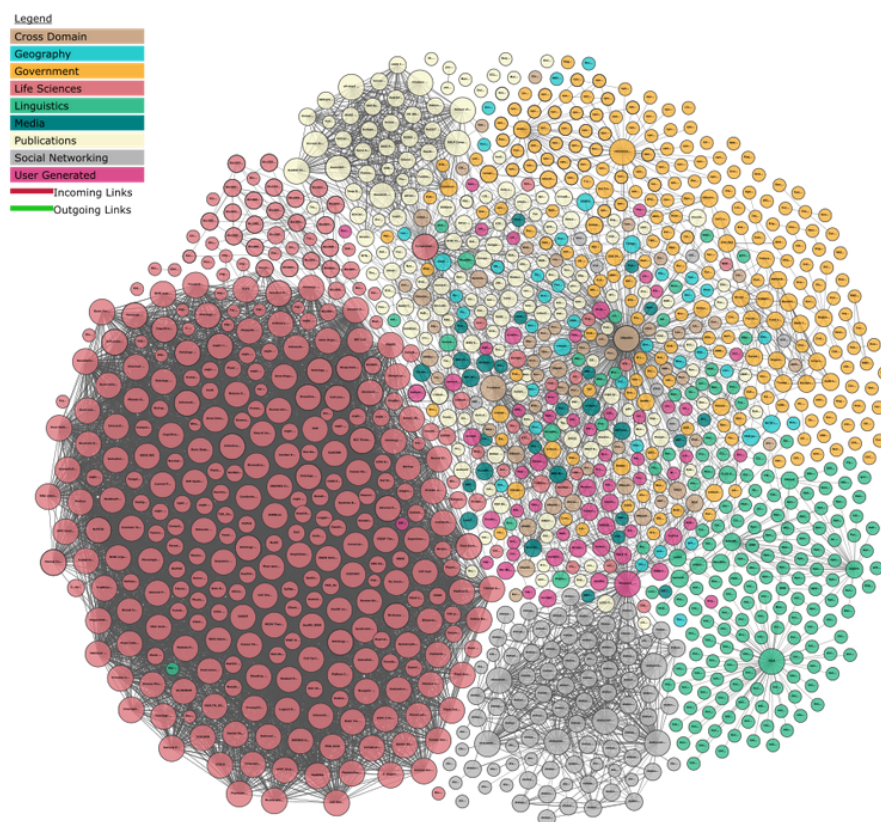


FIGURE E.4 – Les données liées en 2017.



## Extrait du code source de la page de la BnF répertoriant les « signets » en format Dublin core

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="fr" lang="fr">
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1"/>
    <link rel="stylesheet" type="text/css" href="styles/signets.css"/>
    <title>Les Signets de la Bibliothèque nationale de France - Accueil</title>
    <meta name="author" content="Bibliothèque nationale de France"/>
    <meta name="description" content="Les Signets de la Bibliothèque
nationale de France - Accueil - Les Signets de la Bibliothèque nationale
de France proposent un recensement encyclopédique commenté
de ressources gratuitement accessibles par Internet, sélectionnées
et mises à jour par les bibliothécaires de la BnF,
en particulier à destination du public
de l'université et de la recherche."/>
    <meta name="DC.Title"
content="Les Signets de la Bibliothèque nationale de France"/>
    <meta name="DC.Title.Alternative" content="Les Signets de la BnF"/>
    <meta name="DC.Creator" content="Bibliothèque nationale de France,
Direction des collections"/>
    <meta name="DC.Subject" scheme="RAMEAU"
content="Bibliothèque nationale de France"/>
    <meta name="DC.Subject" scheme="RAMEAU"
content="Signets (Web)"/>
    <meta name="DC.Subject" scheme="RAMEAU"
content="Répertoires de liens (Web)"/>
    <meta name="DC.Subject.Classification"
scheme="DDC" content="025.52"/>
    <meta name="DC.Subject.Classification"
scheme="DDC" content="025.56"/>
    <meta name="DC.Description" content="Les Signets de la Bibliothèque
nationale de France proposent un recensement encyclopédique commenté
de ressources gratuitement accessibles par Internet, sélectionnées et mises à jour
par les bibliothécaires de la BnF, en particulier à destination du public
de l'université et de la recherche."/>
```

```
<meta name="DC.Publisher" content="Bibliothèque nationale de France"/>
<meta name="DC.Contributor" content="Bibliothèque nationale de France,
Service de coordination internet"/>
<meta name="DC.Date.Issued" content="2004-09-30"/>
<meta name="DC.Date.Created" content="2004-07-01"/>
<meta name="DC.Type" content="text"/>
<meta name="DC.Format" content="text/html"/>
<meta name="DC.Identifier" content="http://signets.bnf.fr"/>
<meta name="DC.Language" scheme="ISO639" content="fr"/>
<meta name="DC.Rights" content="Bibliothèque nationale de France"/>
<link rel="shortcut icon" type="image/x-icon" href="icono/favicon.ico"/>
<link rel="icon" type="image/gif" href="icono/favicon.gif"/>
</head>
```

# Les critères de qualité définis par Gherasim *et al.* [2012]

Cette annexe reprend le cadre des problèmes de qualité définis dans Gherasim *et al.* [2012]. Nous fournissons des résumés des définitions lisibles dans l'article.

Le cadre			La définition du problème	Réponse de la méthode LOVMI
Logique	Erreurs	Inconsistance logique	Contradiction logique	E1 : Raisonneur
		Ontologie inadaptée	Modélisation ontologique inadaptée au modèle	E2 : OOPS!; E5 : Entretien avec les acteurs
		Ontologie incomplète	Modélisation ontologique incomplète par rapport au modèle	E2 : OOPS!; E5 : Entretien avec les acteurs
		Raisonnement incorrect	Inférence logique déduite de l'ontologie, mais qui ne correspond pas au modèle	
		Raisonnement incomplet	Inférence logique vraie pour le modèle, mais qui ne peut être déduite dans l'ontologie	E2 : OOPS!
	Situations inadaptées	Équivalence logique de deux artefacts	Deux artefacts (concepts, relations ou instances) sont équivalents, mais modélisés distinctement	E2 : OOPS!; E5 : Entretien avec les acteurs
		Artefacts symétriques, logiquement indissociables	Deux artefacts (concepts, relations ou instances) sont modélisés distinctement, mais sans preuve logique. Par exemple, un concept non connecté à l'ontologie	E2 : OOPS!; E5 : Entretien avec les acteurs
		Artefact OR	Un artefact A équivalent à une disjonction de classes. Par exemple, le concept Humain est une personne femme ou une personne homme, et s'il existe une instance commune aux deux classes, alors Humain ne correspond plus à sa définition logique	E1 : Raisonneur; E2 : OOPS!
		Artefact AND	Un artefact A équivalent à une conjonction de classes. Par exemple, le concept Hermaphrodite est une personne femelle et une personne mâle, ces deux concepts doivent donc partager des propriétés ou des instances communes	E1 : Raisonneur; E2 : OOPS!
		Non satisfaite	Un artefact non satisfait ne peut contenir d'instance, il n'est pas vraie pour le modèle décrit	E1 : Raisonneur
		Complexité élevée de la tâche de raisonnement	Quand quelque chose est exprimé de manière à compliquer le raisonnement, alors qu'il existe une façon plus simple de le dire.	
		Ontologie non minimale	L'ontologie contient des informations non pertinentes. Par exemple, A est un B, B est un C et A est C. Cette dernière affirmation peut-être dérivée des deux premières. Autre exemple, si un concept A est défini dans l'ontologie, alors qu'il ne l'est pas dans le modèle.	E2 : OOPS!; E5 : Entretien avec les acteurs

		<b>Le cadre</b>	<b>La définition du problème</b>	<b>Réponse de la méthode LOVMI</b>
<b>Social</b>	<b>Erreurs</b>	Contradiction sociale	Les acteurs perçoivent l'ontologie comme contradictoire à ce qu'elle définit	E2 : OOPS!; E5 : Entretien avec les acteurs
		Perception des erreurs de conception	Les acteurs perçoivent des erreurs de conception telles que la modélisation d'instances en concepts	E2 : OOPS!; E5 : Entretien avec les acteurs
		Socialement vide de sens	Les acteurs ne peuvent comprendre le sens de l'ontologie ou des artefacts qu'elle contient, tel qu'un nom de concept artificiel 'XXXI'	E5 : Entretien avec les acteurs
		Socialement incomplet	Les acteurs perçoivent des manques dans la modélisation ontologique	E2 : OOPS!; E5 : Entretien avec les acteurs
	<b>Situations inadaptées</b>	Manque ou absence d'explications textuelles	Compréhension difficile de l'ontologie par manque d'explication en langage naturel	E2 : OOPS!; E3 : Requêtes SPARQL; E5 : Entretien avec les acteurs
		Artefact potentiellement équivalent	Les acteurs identifient des équivalences au sein de l'ontologie. Par exemple, des concepts ou relations synonymes, ou des concepts dénotés par le même label.	E2 : OOPS!; E3 : Requêtes SPARQL; E4 : Choix du label; E5 : Entretien avec les acteurs
		Artefact indissociable	Les acteurs ne peuvent différencier deux artefacts modélisés distinctement. Par exemple, un label polysémique assigné à deux concepts différents sans que l'on puisse identifier la différence de sens.	E3 : Requêtes SPARQL; E4 : Choix du label; E5 : Entretien avec les acteurs
		Labels polysémiques	Artefacts qui peuvent être identifiés comme l'union ou l'intersection de concepts plutôt que deux concepts distincts	E2 : OOPS!; E5 : Entretien avec les acteurs
		Planéité de l'ontologie ou non modularité	Ontologie présentée comme un ensemble d'artefacts sans structure additionnelle	E5 : Entretien avec les acteurs
		Formalisation non standard	Utilisation d'une logique ou d'une théorie très spécifique qui gêne la compréhension et l'utilisation de l'ontologie	E2 : OOPS!; E5 : Entretien avec les acteurs; E6 : Adéquation sémantique du modèle pour une application
		Version de l'ontologie inadaptée et non certifiée	La version de l'ontologie demande des efforts particuliers pour être comprise et cela gêne la compréhension et l'utilisation de l'ontologie	E5 : Entretien avec les acteurs; E6 : Adéquation sémantique du modèle pour une application
		Artefacts inutiles	L'ontologie contient des artefacts inutiles	E2 : OOPS!; E5 : Entretien avec les acteurs

# Anonymisation des comptes rendus d'hospitalisation

Nous présentons dans cet annexe, l'anonymisation qui a été réalisée sur notre corpus pour en permettre son exploitation.

## H.1 Le respect de la confidentialité pour l'exploitation de données médicales

Nous présentons dans cette section le travail de recherche que nous avons réalisé pour garantir l'anonymisation de notre corpus.

### H.1.1 L'anonymisation de dossiers médicaux

Les données dites « sensibles », dont font parties les données de santé, sont encadrées pas la loi 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, dont une version consolidée est paru au 27 août 2011<sup>1</sup>. Toutefois, aucune recherche dans le domaine de la santé ne peut être mise en œuvre sans l'accord préalable de la CNIL. Cette dernière rappelle également dans son guide pratique « Informatique et libertés » que « la loi « Informatique et Libertés » est applicable dès lors qu'il existe un traitement automatisé ou un fichier manuel, c'est-à-dire un fichier informatique ou un fichier papier contenant des informations personnelles relatives à des personnes physiques ». Un ouvrage parut récemment reprend en détail les articles relatifs à la loi Informatique et libertés [Matta-tia, 2013] et rappelle également que l'Union Européenne n'est pas en reste car elle s'est elle aussi dotée, en 1995, d'une directive commune à l'ensemble des pays membres : la directive Européenne 95/46/CE<sup>2</sup>. Cette directive, largement inspirée des textes français, est toutefois moins restrictive que la loi « Informatique et Libertés » et n'impose pas de disposer d'une organisation de contrôle, telle que la CNIL. Dans ces textes est également précisé que le chaînage de fichiers nominatifs, pour la recherche médicale requiert que

---

1. <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=LEGITEXT000006068624&dateTexte=vig>  
2. <http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT00000697074>



l'information soit rendue anonyme. En somme, tout traitement sur des données médicales doit répondre aux contraintes définies dans ces textes. L'informatique « ne [devant] porter atteinte ni à l'identité humaine, ni aux droits de l'homme, ni à la vie privée, ni aux libertés individuelles ou publiques »<sup>3</sup>.

### Les données nominatives :

Les données nominatives font références à toutes les données permettant d'identifier une personne directement (entre autres par le nom, nom de jeune fille, prénom, par la date de naissance) et indirectement (par le numéro de sécurité sociale, numéro de téléphone, l'âge ou encore l'adresse). Il est précisé, selon la CNIL, que les données directement nominatives doivent être cryptées avec des clefs de cryptage d'au moins 56 bits. Alors que les données indirectement nominatives nécessitent un cryptage avec des clefs d'au moins 40 bits [Quantin *et al.*, 2013].

En France, aucun texte ne définit exhaustivement les données nominatives. L'appréciation est laissée au bon jugement des personnes qui manipulent ces données, la CNIL veillant en amont au respect de l'anonymat des personnes. Ce fait n'est cependant pas le même dans tous les pays. Aux Etats-Unis, la loi HIPAA<sup>4</sup> (Health Insurance Portability and Accountability Act) permet de trancher par une liste officielle constituée de 18 identifiants (PHI<sup>5</sup>) à supprimer, dont : nom, prénom, lieu, date, âge (si + de 89), téléphone, télécopie, adresse courriel, numéro de sécurité social, enregistrement médical, complémentaire santé, compte bancaire, carte d'identité, et de permis de conduire, références sur le véhicule, URLs, adresse IP, numéro de série ou identifiant d'appareil implémenté ou identifiant biométrique [Grouin *et al.*, 2009].

### H.1.2 Les méthodes d'anonymisation

Les études de Meystre *et al.* [2010] et de Uzuner *et al.* [2007]

Les recherches dans le domaine médical étant croissantes, les besoins en anonymisation de données médicales poussent au développement de méthodes automatiques d'anonymisation. Une étude d'évaluation [Meystre *et al.*, 2010], sur les outils disponibles pour l'anglais, a été réalisée en analysant près de 200 publications. Une liste exhaustive permet d'identifier les résultats des outils développés et de discuter des méthodes utilisées. En outre, les auteurs remarquent que les méthodes basées sur des dictionnaires ont de meilleurs résultats, mais sont par contre difficilement généralisables. Uzuner *et al.* [2007] ont réalisé une autre étude d'évaluation pour l'anglais. Sept équipes de chercheurs étaient en compétition pour développer le meilleur outil d'anonymisation. L'évaluation conclue que les systèmes d'apprentissages basés sur des modèles de règles, donnent de meilleurs résultats que les autres systèmes, tel par exemple ceux basés uniquement sur de l'apprentissage. Ils observent également que l'ambiguïté des termes détériore grandement les systèmes. Enfin, les auteurs pointent du doigt la nécessité de développer des systèmes robustes aptes à faire face à l'hétérogénéité des données provenant de sources

3. Article 1 de la loi 78-17 du 6 janvier 1978.

4. <http://www.hhs.gov/ocr/privacy/hipaa/administrative/privacyrule/index.html>

5. protected health information

différentes ou sous des formats divers. Ainsi que l'importance de délimiter les utilisations possibles de ces outils, selon les niveaux de sécurité demandés.

### Le projet ALADIN-DTH

ALADIN-DTH<sup>6</sup> (Assistant de lutte automatisée de détection des infections nosocomiales à partir de documents textuels hospitaliers) [Gicquel *et al.*, 2012] visait au développement d'un outil pour anonymiser les comptes rendus d'hospitalisation (CRH) nécessaires au projet. L'outil se base sur XEROX INCREMENTAL PARSER<sup>7</sup>, un système de reconnaissance d'entités nommées développé pour le français. Les chercheurs ont adapté cet outil à leur corpus du domaine médical en procédant à : « un enrichissement lexical, une adaptation de la désambiguïsation des parties du discours, une adaptation au format des textes analysés et un traitement spécifique des dates ». Actuellement, l'outil ne présente pas encore de résultats satisfaisants permettant d'en envisager un usage automatique, mais de nouveaux développements sont en cours. De plus, suite aux premiers essais, les chercheurs concluent à une nette réduction du temps d'anonymisation manuelle.

### Le logiciel MEDINA

Un autre outil, présentant de meilleurs résultats, a été développé pour le français dans le cadre de la thèse de Grouin [2013], le logiciel MEDINA [Grouin *et al.*, 2009]. Cet outil, spécifiquement développé pour l'anonymisation de données médicales permet d'anonymiser des dossiers avec un taux de précision d'environ 91% et un taux de rappel d'environ 85%. Ce logiciel, inspiré en grande partie de l'anonymiseur DE\_IT [Neamatullah *et al.*, 2008], réalisé par des équipes du MIT<sup>8</sup> utilise des expressions régulières pour anonymiser les entités numériques et des dictionnaires et listes d'entités nommées pour anonymiser les noms propres. L'anonymisation se réalise en deux étapes. Une première anonymisation est réalisée sur le corpus, afin d'identifier les noms et prénoms via les dictionnaires. Ensuite, une deuxième anonymisation étudie le voisinage des termes déjà anonymisés et repère un mot à la droite ou à la gauche, qui commence par une majuscule et qui n'est pas dans le dictionnaire de noms communs. Le logiciel en déduit qu'il s'agit d'un nom de famille ou d'un prénom. À noter que cet outil, tout en garantissant l'anonymisation des dates, permet de conserver la chronologie des événements médicaux, ainsi que les intervalles de temps, en établissant un décalage temporel.

### H.1.3 Les méthodes de chiffrage et de chaînage des données

Le chaînage des données de santé est une problématique majeure des études épidémiologiques. En effet, le but du chiffrage des données est qu'elles soient rendues invisibles à toute personne et à tout système informatique. Cependant, certaines études nécessitent de savoir quelles données vont ensemble. Par exemple, si le dossier d'un patient est divisé en différents fichiers, il est parfois essentiel de savoir que les informations présentes dans ces différents fichiers font références au même patient.

---

6. <http://www.aladin-project.eu/>

7. <http://open.xerox.com/Services/XIPParser>

8. Massachusetts Institute of Technology : <http://www.mit.edu/>

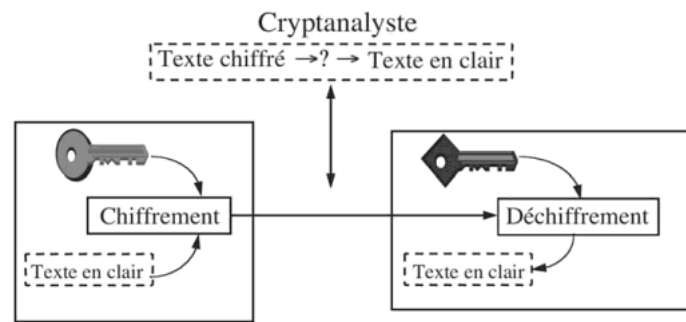


FIGURE H.1 – Association de deux clefs privées servant à crypter et décrypter un texte par [Quantin *et al.*, 2013].

### Les algorithmes cryptographiques :

Les algorithmes permettant de crypter les informations sont nombreux et permettent grâce à des clefs, de déchiffrer des messages échangés. Nous trouvons des systèmes de cryptage dits à clef secrète, lorsqu’une seule clef sert à la fois au cryptage et au décryptage (voir figure H.1). C’est le cas par exemple de l’algorithme DATA ENCRYPTION STANDARD (DES) [Mercier, 1996]. Ce mode de cryptage très populaire a d’ailleurs été retenu en janvier 1977 pour servir à toutes les organisations fédérales aux États-Unis. Mais comme le souligne Mercier [1996], il montre vite ses limites, car il s’adresse à un public large. Il faut donc distribuer suffisamment de clefs pour que tous les couples émetteurs-récepteurs en possèdent. Donc si une société souhaite échanger des informations cryptées avec un millier de ses clients, il lui faudra garder secrète un millier de clefs tout en autorisant des personnes « sûres » à y accéder. Nous voyons donc rapidement apparaître les problèmes de partage de clef et de sécurité qu’impliquent ces systèmes.

Un autre système de cryptage est dit à clef révélée. Ce type de système repose sur l’utilisation de deux clés. La première est publique et tout le monde peut l’utiliser pour envoyer un message chiffré à un destinataire donné. La seconde est privée, elle est connue uniquement de ce destinataire (et ne nécessite donc pas de stockage, comme c’est le cas avec les systèmes à clef secrète) et elle seule peut permettre de décrypter le message [Quantin *et al.*, 2013]. L’algorithme à clé révélée le plus connu est l’algorithme RIVEST SHAMIR ET ADLEMAN (RSA).

### Les algorithmes de hachage :

Le but d’un algorithme de hachage est qu’il ne soit pas réversible. Une fois l’information cryptée, il ne doit plus exister aucune méthode permettant de décrypter le message. Il ne s’agit donc pas uniquement de rendre l’information illisible, mais également de la rendre indéchiffrable. De nombreux algorithmes de hachage irréversibles existent, tels que les SHA ou MD5. El Kalam *et al.* [2004] a ainsi développé un système afin de chaîner et crypter des données. Chaque patient est identifié avec un identifiant unique ne pouvant être décrypté. Cette suite de caractères est ce qu’on appelle un hash et est généré de différentes manières suivant les algorithmes de hachage utilisés. Dans le système inventé par El Kalam *et al.* [2004], le hash de chaque patient est généré grâce à un identifiant

patient unique, stocké dans la carte vitale et associé à un identifiant projet.

## H.2 Anonymisation du corpus avec l'aide du logiciel MEDINA

Dans cette section, nous présentons la mise en œuvre de l'anonymisation de notre corpus et du chaînage des informations liées aux données des patients. Un accès aux CRH à anonymiser a été possible durant le mois d'octobre 2013 au SHU de Sainte-Anne. À cette occasion, les 8 700 CRH ont pu être anonymisés avec l'aide du logiciel MEDINA (H.1.2), puis ensuite validés par l'équipe responsable du projet à l'hôpital. Le logiciel MEDINA a été développé dans le cadre de la thèse de [Grouin \[2013\]](#), il est distribué librement. MEDINA fonctionne avec cinq modules, que l'on peut lancer indépendamment les uns des autres sur le corpus. Afin de pouvoir chaîner les données anonymisées, nous nous servons de l'indépendance des modules et anonymisons les CRH étape par étape. Ceci après avoir formaté et nettoyé en partie le corpus pour que tous les fichiers soient en format texte (.txt). La méthodologie est schématisée figure [H.2](#).

### H.2.1 Annotation des entités nommées - étape 1

Le premier module permet d'annoter les entités à anonymiser. Ici les noms, prénoms, adresses, dates, numéros de téléphone et de sécurité sociale. Cette étape est réalisée en deux passages sur les textes. Un premier passage permet d'annoter les noms et prénoms contenus dans les dictionnaires ou ceux précédés d'un déclencheur (Madame, Monsieur, Docteur, etc.). Un deuxième passage est ensuite réalisé sur le voisinage des termes déjà annotés. Cela permet d'annoter un nom ou un prénom qui n'aurait pas été détecté au premier passage. Prenons l'exemple « Mme Pappenheim Bertha ». Si le premier passage permet d'annoter « Pappenheim » en nom grâce au déclencheur « Mme », mais que le prénom « Bertha » n'est pas dans le dictionnaire, celui ci ne sera pas annoté. Le deuxième passage va permettre d'identifier que le nom « Pappenheim » est suivi d'un autre nom propre « Bertha ». Le système va ainsi en déduire que « Bertha » est à annoter en prénom. À cette étape, les fichiers sont balisés de la manière suivante :

```
Adresse du centre hospitalier
Compte-Rendu d'Hospitalisation de
<nom>Pappenheim</nom> <prenom>Bertha</prenom>
Né le <date>27/02/1859</date>
Numéro de téléphone : <numero>XX XX XX XX XX</numero>
Numéro de sécurité social : <numero>X XX XX XX XXX XXX XX</numero>
Adresse : <adresse>106 Rue de la Santé, 75014 Paris</adresse>
Date d'entrée : <date>01 janvier 1880</date>
Date de sortie : <date>01 mars 1880</date> Suivi par : <nom>Breuer</nom>
```

### H.2.2 Nettoyage des fichiers - étape 2

Nous avons constaté à cette étape un très mauvais balisage des numéros et également des noms isolés. De nombreux fichiers sont rédigés de la manière de l'exemple précédant, sans déclencheur devant le nom du médecin ou même du patient et sans typographie

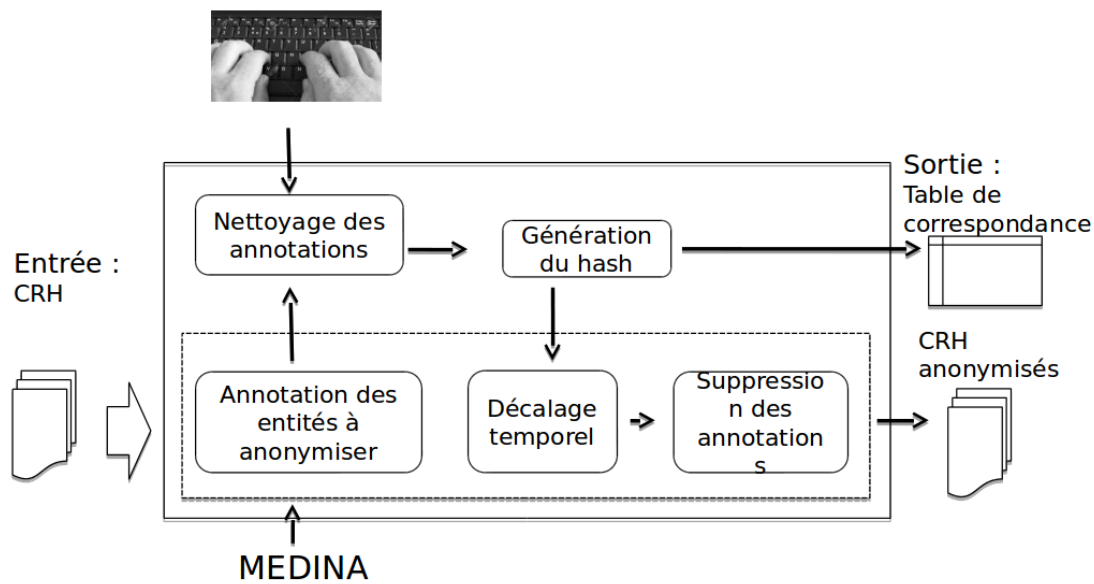


FIGURE H.2 – Procédure d'anonymisation mise en place dans le cadre de notre projet.

unifiée. Cela rend difficile la détection de ces entités et pose problème pour la suite des traitements. Nous avons à cette étape réalisé un nettoyage de l'en-tête des fichiers. Nous avons supprimé à l'aide d'expressions régulières les lignes qui contenaient les numéros de téléphone et de sécurité sociale, ainsi que celles qui contenaient le nom du ou des médecins suivant le patient. Nous avons également corrigé les annotations des noms, prénoms et dates de naissance des patients, qui étaient parfois manquantes. Cette étape a nécessité un traitement manuel dont le rendement était approximativement de 100 CRH traités par heure. À la fin de ce traitement les en-têtes des CRH étaient comme suit :

```

Monsieur/Madame <nom>Pappenheim</nom> <prenom>Bertha</prenom>
Né le <date>27/02/1859</date>
Date d'entrée : <date>01 janvier 1880</date>
Date de sortie : <date>01 mars 1880</date>

```

### H.2.3 Génération d'un hash unique à chaque patient - étape 3

Le hachage des données consiste au cryptage des données sensibles, afin de ne pouvoir identifier un patient. Nous pouvons vouloir supprimer ces données ou bien conserver un chaînage de l'information anonymisée. En effet, les 8700 CRH ne font pas référence à 8700 patients différents, car toute nouvelle hospitalisation entraîne la rédaction d'un nouveau CRH. Le chaînage des CRH n'est pas pertinent pour le développement de l'ontologie. Cependant, il peut l'être pour des études qui nécessitent de croiser des données, nous l'avons donc réalisé afin que les CRH puissent être utilisés dans d'autres études. La génération d'un hash unique à chaque patient s'est faite de manière automatique à l'aide d'un script perl. Une fonction de hachage, l'algorithme md5, a été appliquée sur chaque

nom, prénom et date de naissance des patients. Le résultat de ce hash est inscrit dans chaque CRH à la place du nom, prénom et de la date du naissance du patient, ainsi que dans une table de correspondance à l'attention du personnel chargé de l'étude à l'hôpital Sainte-Anne. Exemple d'entrée de la table de correspondance :

Nom,Prénom,Date de naissance,Hash,Numero du CRH  
Bernard,Paul,05/03/1967,bhdzjaklafhdsjfbrjhe,345

### **H.2.4 Décalage temporel - étape 4**

Le logiciel MEDINA dispose d'un module pour anonymiser les dates, tout en conservant la chronologie des événements relatés dans les CRH. Pour cela il effectue un décalage temporel équivalent sur toutes les dates de chaque fichier. Et ce décalage est généré de manière aléatoire sur les fichiers.

Date avant le décalage temporel :

Date d'entrée : <date>01 janvier 1880</date>  
Date de sortie : <date>01 mars 1880</date>  
Date après le décalage temporel (ici décalage +10 jours, +3 mois, -7 ans) :  
Date d'entrée : <date>11 avril 1873</date>  
Date de sortie : <date>11 juin 1873</date>

### **H.2.5 Suppression du contenu des balises ou de toutes traces d'entités - étape 5**

Les derniers modules permettent d'effacer les traces des entités nommées. Au choix, nous pouvons remplacer les annotations par un blanc ou par une balise fermante. Dans notre cas, nous remplaçons les entités par des blancs. En effet, la phase suivante étant l'extraction de termes candidats, il nous est inutile - voire même gênant pour la suite des traitements - de conserver des informations de balisage.

### **H.2.6 Résultats et discussions de l'anonymisation avec MEDINA**

L'outil MEDINA a permis d'aider à l'anonymisation des CRH. Nous avons défini sept identifiants à anonymiser : noms et prénoms (incluant les noms et prénoms des praticiens et de toutes personnes citées dans les CRH), adresses y compris codes postaux et villes, dates, noms d'hôpitaux, numéros de sécurité social et de téléphone. Faute de temps, nous n'avons pas pu mesurer statistiquement les résultats obtenus avec le logiciel MEDINA (nous disposions d'un accès aux CRH non anonymisés limité dans le temps et restreint à l'enceinte de l'hôpital).

La reconnaissance des noms et prénoms non enregistrés dans le dictionnaire de MEDINA n'a pas donné de bons résultats au premier test (dû généralement à l'absence de déclencheur). Les fichiers étaient enregistrés aux noms et prénoms des patients. Nous avons donc réalisé une extraction de ces entités à l'aide d'une expression régulière. Nous avons ensuite ajouté ces noms et prénoms au dictionnaire de MEDINA, ainsi que la liste des noms des praticiens de Sainte-Anne. Cela a permis d'améliorer les performances de reconnaissance et d'annotation de noms et prénoms des patients et des praticiens.

Enfin, de nombreuses dates restent non balisées. Les expressions régulières visant à leur détection doivent donc être améliorées. Nous ne pouvons tenir compte de la chronologie si nous ne sommes pas certains que toutes les dates présentent dans les CRH ont été traitées de la même manière. En d'autres termes, il suffit que dans un CRH une date n'ait pas été reconnue pour que celle-ci échappe au décalage temporel et soit donc inexploitable, pour la suite de notre étude. De plus, une date non décalée ne respecte pas la législation concernant l'anonymisation des données « sensibles » et pose donc problème d'un point de vue éthique et administratif.

### **H.3 Synthèse**

L'anonymisation, qui est une tâche parallèle à nos travaux, a demandé un important travail préparatif et une étude attentive des législations en vigueur, pour la recherche appliquée à des données médicales. Le logiciel MEDINA nous a grandement aidé pour la réalisation de cette tâche.

## Table de correspondance entre les étiquettes de TREETAGGER et de MELT

MELT	TreeTagger	MELT	TreeTagger
cc, cs	kon	npp	nam
adj, adjwh	adj	v	ver :pres
adv, advwh	adv	P+D	PRP :det
clo, clr, cls	pro :per	PONCT (, ; :() +)	PUN
p	prp	PUN :cit ( »)	PUN
prorel	pro :rel	.,..., ?, !	SENT
vpp	ver :pper	ma,ta,sa,...+DET	DET :POS
vs	ver :subi	DET	DET :ART
vpr	ver :ppre	09+ ADJ	NUM
vinf	ver :infi	PRO, P+PRO	PRO
vimp	ver :impe	i	int
nc	nom		





## Extrait du fichier xml résultat de l'extraction avec YATeA

```

<TERM_CANDIDATE MNP_STATUS="1">
----<ID>term3</ID>
----<FORM>trouble de la personnalité</FORM>
----<LEMMA>trouble de le personnalité</LEMMA>
----<MORPHOSYNTACTIC_FEATURES>
-----<SYNTACTIC_CATEGORY>NOM de DET:ART NOM</SYNTACTIC_CATEGORY>
----</MORPHOSYNTACTIC_FEATURES>
----<HEAD>term4</HEAD>
----<NUMBER_OCCURRENCES>1</NUMBER_OCCURRENCES>
----<LIST_OCCURRENCES>
-----<OCCURRENCE>
-----<ID>occ3</ID>
-----<MNP>1</MNP>
-----<DOC>0</DOC>
-----<SENTENCE>0</SENTENCE>
-----<START_POSITION>97</START_POSITION>
-----<END_POSITION>123</END_POSITION>
----</LIST_OCCURRENCES>
----<TERM_CONFIDENCE>0.5</TERM_CONFIDENCE>
----<TERM_WEIGHTS>
-----<WEIGHT name="DDW">0</WEIGHT>
----</TERM_WEIGHTS>
----<LOG_INFORMATION>YaTeA</LOG_INFORMATION>
----<SYNTACTIC_ANALYSIS>
-----<HEAD>term4</HEAD>
-----<MODIFIER POSITION="AFTER">term5</MODIFIER>
-----<PREP>
-----de
-----</PREP>
-----<DETERMINER>
-----la
-----</DETERMINER>
----</SYNTACTIC_ANALYSIS>
</TERM_CANDIDATE>

```



## Extrait du fichier xml résultat de l'extraction avec BIOTEX

```

<Terms>
--<Term id="1">
----<name_term>
-----trouble de l'adaptation avec humeur anxiodépressive
----</name_term>
----<is_true_term validated_by="">0</is_true_term>
----<score>3.0</score>
--</Term>
--<Term id="2">
----<name_term>
-----trouble de l'adaptation avec humeur
----</name_term>
----<is_true_term validated_by="">0</is_true_term>
----<score>2.8074</score>
--</Term>
--<Term id="3">
----<name_term>
-----adaptation avec humeur anxiodépressive
----</name_term>
----<is_true_term validated_by="">0</is_true_term>
----<score>2.3219</score>
--</Term>
--<Term id="4">
----<name_term>
-----trouble de l'adaptation
----</name_term>
----<is_true_term validated_by="MeSH">1</is_true_term>
----<score>2.3219</score>
--</Term>
--<Term id="5">
----<name_term>
-----trouble de la personnalité
----</name_term>
----<is_true_term validated_by="MeSH">1</is_true_term>
----<score>2.3219</score>
--</Term>
--<Term id="6">
----<name_term>

```

```
-----adaptation avec humeur
----</name_term>
----<is_true_term validated_by="">0</is_true_term>
----<score>2.0</score>
--</Term>
--<Term id="7">
----<name_term>
-----humeur anxiodépressive
----</name_term>
----<is_true_term validated_by="">0</is_true_term>
----<score>1.585</score>
--</Term>
--</Terms>
```

## Arborescence conceptuelle de l'ontologie sur les facteurs sociaux et environnementaux

La figure [L.1](#) présente les premiers niveaux de la hiérarchie conceptuelle des branches : *concept de changement*, *concept de comportement* et *concept de condition de vie* à gauche. Ceux des branches *concept de caractéristique* et *concept de droit et justice française* à droite.

La figure [L.2](#) présente le premier niveau de la hiérarchie conceptuelle des branches *vie sociale*, *être vivant* (et *être humain*), *groupe* (et *groupe primaire*, *groupe secondaire*, *groupe d'individus*) à gauche. Premier niveau de la hiérarchie conceptuelle de la branche *individu*, *individu selon qualificatif*, *individu selon relation* *individu selon relation familiale*, *individu selon relation juridique* *protecteur juridique* à droite.

La figure [L.3](#) présente les premiers niveaux de la hiérarchie conceptuelle des branches *concept d'éducation*, *concept événement*, *concept de lieu* et à droite. Ceux des branches *concept relationnel*, *concept sentiment*, ainsi que *état émotionnel négatif*, *positif* et *variable* à droite.

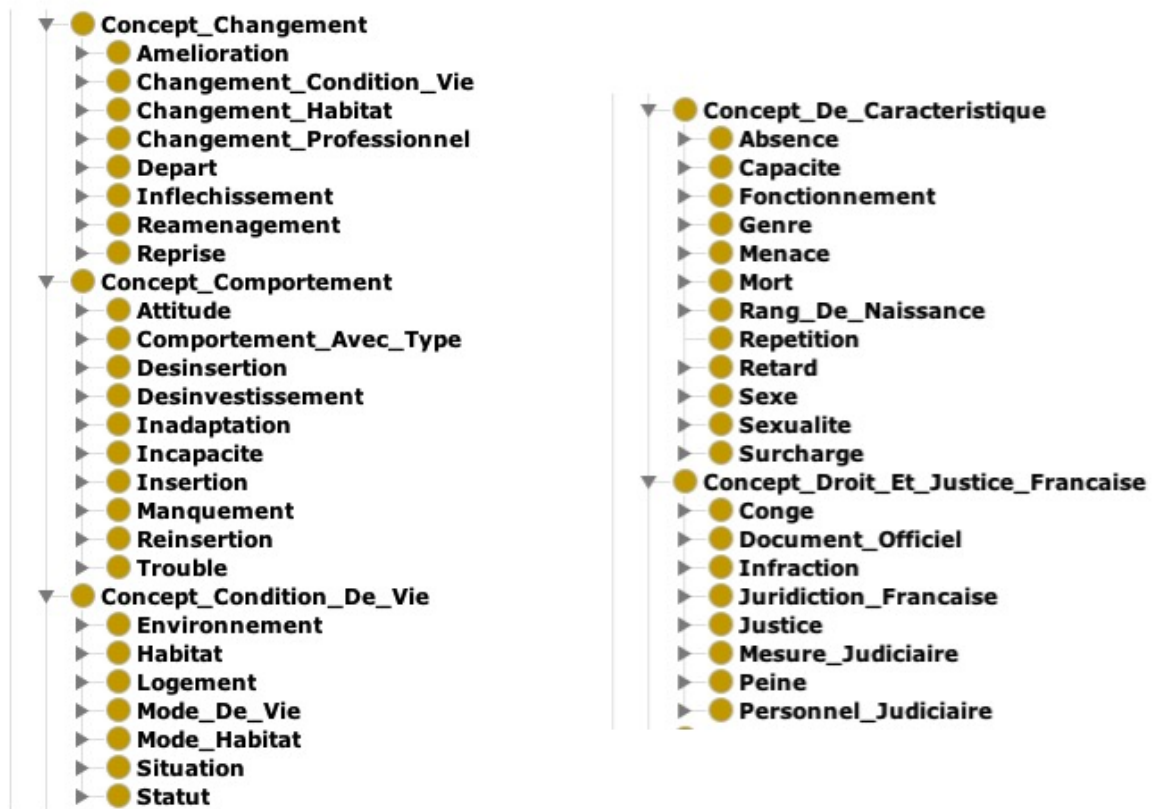


FIGURE L.1 – Extrait 2 de la hiérarchie conceptuelle de l'ontologie.



FIGURE L.2 – Extrait de la hiérarchie conceptuelle de l'ontologie.





FIGURE L.3 – Extrait de la hiérarchie conceptuelle de l'ontologie.

# Annexe M

## Arborescence conceptuelle construite à partir des termes extraits des comptes rendus d'hospitalisation



FIGURE M.1 – Premier niveau de hiérarchie conceptuelle de la branche « Acte médical » à gauche, et des branches « Diagnostic » et « Parcours de soins » à droite.



FIGURE M.2 – Premier niveau de hiérarchie conceptuelle de la branche « Etat » à gauche, et de la branche « Trouble » à droite.



FIGURE M.3 – Premier niveau de hiérarchie conceptuelle de la branche « Observation de la condition du patient » à gauche, de la branche « Observation de l'évolution de la maladie » au centre, et de la branche « Observation médicale générale » à droite.



## RÉSUMÉ

La psychiatrie est une spécialité médicale qui vise à fournir un diagnostic et à traiter des troubles mentaux. Malgré des classifications internationalement reconnues, la catégorisation des patients selon des critères diagnostiques reste problématique. Les catégories actuelles peinent à prendre en compte l'hétérogénéité interindividuelle, les difficultés de délimitations des syndromes et l'influence sur les symptômes de nombreux facteurs dans l'histoire individuelle ou dans l'environnement. La recherche en psychiatrie nécessite une amélioration de la description des comportements, des syndromes ou des dysfonctionnements associés aux troubles psychiatriques. À cette fin, nous proposons ONTOPSYCHIA, une ontologie pour la psychiatrie, divisée en deux modules : les facteurs sociaux et environnementaux des troubles mentaux, et les troubles mentaux. L'utilisation d'ONTOPSYCHIA associée à des outils dédiés permettra la prise en compte des facteurs sociaux et environnementaux, la représentation de la comorbidité et une proposition de consensus autour des catégories descriptives des troubles psychiatriques. Dans un premier temps, nous avons développé les deux modules ontologiques selon deux méthodes différentes. La première propose une analyse de comptes rendus d'hospitalisation, tandis que la deuxième propose un alignement de différentes classifications psychiatriques, pour répondre au besoin de consensus. Dans un deuxième temps, nous avons développé un cadre méthodologique pour valider la structure et la sémantique des ontologies.

**Mots-clés :** informatique médicale, ontologie, ingénierie des connaissances, psychiatrie, classification médicale, validation d'ontologies.

---

## ABSTRACT

Psychiatry is a medical speciality that aims at providing diagnosis and treating mental disorders. Despite internationally acknowledged criteria leading to diagnostic categories, most psychiatric disorders are syndromes with common symptoms or dimensions between these diagnostic categories. In addition, the analysis of the prevalence and incidence of social and environmental risk factors of diseases is crucial to understand and treat them and might have significant impacts on policy decisions (therapeutic as well as the length or the cost of hospitalisation). This overlap between diagnoses and the heterogeneity within the defined diagnoses stresses the need to improve our capability to detect, to quantify the behaviour and to model the symptoms and the social and environmental risk factors associated to psychiatric disorders. To that end, we propose ONTOPSYCHIA, an ontology for psychiatry, divided in two modules : social and environmental factors of mental disorders and mental disorders. ONTOPSYCHIA associated with dedicated tools will help to perform semantic research in Patient Discharges Summaries (PDS), to represent comorbidity, to reach a consensus on descriptive categories of mental disorders. In a first step, we developed two ontological modules using two different methods. The first proposes an analysis of PDS, while the second proposes an alignment of psychiatric classifications to meet the need for consensus. In a second step, we have developed a methodological framework to validate the structure and semantics of ontologies.

**Mots-clés :** medical informatics, ontology, knowledge engineering, psychiatry, medical classification, ontology validation.

