Contributions à l'Ingénierie des Connaissances : Construction et Validation d'Ontologie et Mesures Sémantique

Mounira Harzallah

Université de Nantes

École doctorale STIM

Sciences et Technologies de l'Information et Mathématiques

Habilitation à Diriger des Recherches (H.D.R.)

Spécialité : informatique

Contributions à l'Ingénierie des Connaissances

Construction et Validation d'Ontologie et Mesures Sémantiques

par

Mounira Harzallah

Soutenue le 13 décembre 2017 devant le jury composé de :

Mme Nathalie Aussenac-Gilles, Directrice de recherche, CNRS/IRIT, Univ. Paul Sabatier, Toulouse (Rapporteur)

M. Jean Charlet, Chargé de mission recherche à l'AP-HP & Inserm, Université Pierre et Marie Curie (Rapporteur)

M. Serge Garlatti, Professeur, Institut Mines Télécom (IMT) Atlantique (Rapporteur)

Mme Bénédicte Le Grand, Professeur, Université Paris 1 Panthéon - Sorbonne, Centre de Recherche en Informatique (CRI) (Examinateur)

Mme Pascale Kuntz, Professeur, École Polytechnique de l'Univ. de Nantes (Examinateur)

Remerciements

Je remercie mes rapporteurs pour leurs retours enrichissants qui m'ont été d'une grande utilité et m'ont permis de préciser encore mes contributions en recherche.

Je remercie également tous les membres du jury qui ont accepté d'évaluer ce travail.

Je remercie Mme Pascale Kuntz, professeur à l'école Polytechnique de l'Université de Nantes pour ses encouragements lors de la rédaction de ce manuscrit.

Je remercie M. Philippe Leray, professeur à l'école Polytechnique de l'Université de Nantes pour ses remarques constructives sur ce manuscrit.

Je remercie M. Fabrice Guillet, professeur à l'école Polytechnique de l'Université de Nantes pour son soutien et ses encouragements.

Je remercie toutes les personnes avec qui j'ai mené des travaux de recherche dans mon équipe DUKe, en France ou à l'étranger. Je remercie particulièrement, M. Guiseppe Berio, professeur à l'Université de Bretagne Sud pour notre collaboration fructueuse, depuis 1998.

Je remercie Mme Muriel Primot, maître de conférence à l'université de Nantes et ma collègue de bureau pour ses relectures de ce manuscrit, pour son soutien et ses encouragements.

A ma fille chérie Ines

A mon mari chéri

Merci pour votre patience

Sommaire

Liste des figures

Liste des tables

Introduction

Les connaissances sont reconnues depuis un certain moment comme le capital de l'économie immatériel. Leur ingénierie (extraction, modélisation, capitalisation, exploitation, *etc.*) est une problématique permanente et omniprésente dans les activités de chacun. Elle devient un des impératifs majeurs de toute analyse de l'existant ou réflexion stratégique au sein d'une organisation. Elle a connu plusieurs mutations en s'adaptant au cours du temps à l'évolution des connaissances et de leur traitement informatique [Aussenac&Gandon, 2013]. Elle a notamment dû prendre en compte une évolution dans le temps des ressources des connaissances (experts, livres, bases de données, réseaux sociaux, tweeters, données, web sémantique, web des données), de leurs formes (implicite, explicite, structurée, semi ou non structurée), de leurs utilisateurs (organisation, apprenant, utilisateur du web), des supports de leur utilisation (livres, bases de données, systèmes experts, systèmes à bases des connaissances, applications du web sémantique), du volume et de la vitesse de multiplication de leurs ressources, des techniques de leur extraction, des langages de leur représentation, *etc*.

L'ontologie a été définie comme une représentation sémantique pertinente d'une conceptualisation explicite et formelle des connaissances d'un domaine. Dans la communauté de l'ingénierie des connaissances, l'aspect formel des ontologies et la possibilité de raisonner avec elles, ont conduit de nombreux chercheurs à s'intéresser aux ontologies pour bien expliciter et représenter les connaissances d'un domaine et pour les utiliser dans plusieurs applications (*e.g.* des applications d'aide à la décision ou de recherche d'information).

Au passage à l'échelle, le challenge d'utilisation des ontologies dans des projets réels s'est vu confronté à plusieurs obstacles parmi lesquels nous pouvons citer : (1) des méthodes de construction peu opérationnelles et non adaptées à la construction d'ontologies volumineuses ; (2) des méthodes de validation plutôt adaptées à certains problèmes logiques dans une ontologie, difficiles à appliquer à des ontologies de grande taille ; (3) la difficulté de bien choisir une ontologie pour une application, parce que cette ontologie doit bien couvrir le domaine de cette application et prendre en compte ses évolutions, et parce qu'il faut y appliquer une mesure sémantique adéquate dont les résultats devraient être et rester proches du jugement humain, même si cette ontologie évolue. Bien sûr, plusieurs mesures sémantiques ont été proposées dans la littérature mais peu de travaux ont étudié l'adéquation d'une mesure sémantique à une ontologie d'une application.

Dans le web sémantique, la prolifération d'ontologies a accentué les obstacles cités précédemment et a conduit à la connexion d'ontologies pour former des ontologies en réseau, avec la difficulté de gérer et d'exploiter des ontologies de différentes généricités et complexités, portant sur des domaines divers. Par ailleurs, pour l'utilisation d'ontologie dans des systèmes ad-hoc, on s'est orienté vers le choix d'une ontologie très générique, couvrant des domaines variés mais ayant un niveau d'expressivité faible, impliquant un spectre de raisonnement limité.

Enfin, dans le web de données, la problématique de la masse de données s'est accentuée en devenant une problématique de flux de données. Le contenu de ce flux pourrait provenir de sources différentes et hétérogènes, être représenté avec des langages différents et parfois être défini avec des métadonnées différentes et bien sûr avec un vocabulaire dont la sémantique est inconnu. On s'est focalisé sur les techniques de fouille de données, d'apprentissage automatique, *etc.*, pour gérer les paramètres de volume des données, de la vitesse de leur changement, de leur véracité… tout en oubliant le paramètre de leur hétérogénéité sémantique [Gruninger&Obrst, 2014].

Nous nous sommes intéressés depuis plus de 13 ans aux problématiques liées aux ontologies, à leur construction, à leur validation et à leur exploitation, en ingénierie des connaissances. Le domaine de recherche de l'ingénierie des connaissances se focalise « sur l'étude des concepts, modèles, méthodes, techniques et outils permettant de modéliser ou d'acquérir les connaissances ; pour le développement des systèmes réalisant ou aidant l'humain à réaliser des tâches ; pour des domaines comme l'acquisition des connaissances à partir des textes, la recherche d'information sur le web… » [Charlet *et al.* 2000], [Aussenac *et al.* 2014].

Nos contributions à ce domaine de recherche sont principalement des modèles sémantiques, des cadres ou des méthodes dont certaines utilisent des techniques de raisonnement logique, de fouille des données ou de traitement automatique des langues. Elles s'organisent autour de trois axes : (1) l'ingénierie des compétences des ressources humaines et son articulation à l'ingénierie des connaissances, (2) la construction et la validation semi-automatiques d'ontologie à partir des textes, et (3) les mesures sémantiques de comparaison d'objets, utilisant une ontologie.

Ingénierie des compétences des ressources humaines. Dans le premier axe, nous nous sommes intéressés à l'acquisition, la modélisation et la gestion des connaissances portant sur les compétences individuelles. La maîtrise des compétences individuelles au sein d'une organisation et de leur gestion s'est révélée capitale dans une période de départ massif à la retraite. L'importance de ce type de connaissances et la maîtrise de son ingénierie sont toujours d'actualité pour plusieurs raisons : (1) leur diversité, leur rôle stratégique dans l'innovation et la performance d'une organisation et le coût élevé de leur gestion ; (2) leur évolution rapide avec l'évolution des métiers ; (3) la disponibilité de ressources volumineuses à partir desquelles elles pourraient être extraites (*e.g.* CV ou traces de l'utilisateur sur le web).

Pour définir le concept de compétence et faciliter son extraction (d'une façon manuelle ou semi-automatique) ou sa gestion, nous avons proposé, dans nos travaux de thèse, un modèle conceptuel pionnier baptisé CRAI (Competency Resource Aspect Individual) et formalisé en théorie des ensembles. A notre arrivée dans l'équipe DUKe (anciennement COD), pour généraliser ce modèle et le positionner par rapport aux modèles existants portant sur les connaissances, nous avons étudié les liens entre les concepts de compétence et de connaissance et proposé le modèle CKIM (Competency and Knowledge Integrated Model) *qui permet d'intégrer la représentation de ces deux concepts.*

Lors du passage à l'échelle et de la considération des connaissances métiers (mémoire d'entreprise), les représentations conceptuelles se sont avérées insuffisantes pour bien expliciter les connaissances, car elles mettent l'accent sur l'intégrité des données et non sur la sémantique des connaissances. En plus, ces représentations sont difficilement évolutives et réutilisables. Enfin, elles ne permettent pas de faire du raisonnement. On s'est donc intéressé à représenter les connaissances sous la forme d'une ontologie.

Dans ce contexte, nous *avons proposé l'ontologie noyau CRAI des compétences basée sur notre modèle CRAI et une architecture intégrante pour l'ingénierie des compétences. Cette architecture, basée sur une ontologie du domaine et notre ontologie noyau CRAI des compétences, inventorie les ressources des connaissances et les techniques de leur ingénierie pour extraire, modéliser et exploiter les compétences. Nous avons déployé cette architecture intégrante, en développant une technique basée sur les règles d'association (une des techniques de fouille de données) pour compléter l'évaluation des compétences acquises par un individu.* Notre ontologie noyau CRAI des compétences a été aussi utilisée dans le projet CommOnCV pour l'aide à l'extraction des compétences à partir des CVs. Nous avons également appliqué une méthode de classification dynamique à des compétences définies à l'aide de notre ontologie noyau CRAI.

Cette architecture a orienté nos travaux de recherche vers les deux autres axes : l'axe 2 qui porte sur les méthodes et techniques d'ingénierie des connaissances pour la conceptualisation et la validation semi-automatiques d'ontologie à partir de textes, l'ontologie et les méthodes et techniques d'ingénierie des connaisses étant deux composants principaux de cette architecture ; et l'axe 3 qui porte sur les mesures sémantiques de comparaison d'objets, une mesure sémantique étant une technique de cette architecture qui pourrait s'appliquer à une ontologie pour aider à accomplir certains processus d'une organisation.

Construction et validation semi-automatiques d'ontologie à partir des textes. Dans l'axe 2, plusieurs travaux ont proposé des méthodes, techniques et outils pour réaliser les différentes étapes du processus de construction d'ontologie. Pour l'étape de conceptualisation d'ontologie, qui est une des étapes de ce processus les plus lourdes et longues à réaliser, les méthodes proposées en définissent ses sous-étapes. Des patrons de conception (ontology design patterns (ODP)) jouant un rôle semblable à celui des patrons développés en génie logiciel ont été proposés pour développer des ontologies réutilisables [Gangemi&Presutti, 2009]. Des principes et règles de conceptualisation ont aussi été définis, par exemple : le principe de réutilisation d'ontologie ou celui de l'utilisation d'une ontologie noyau [Smith, 2007], [Burita *et al.* 2012] ou les règles d'OntoClean [Guarino&Welty, 2002] pour cadrer la conceptualisation. Cependant, cette étape reste complexe à réaliser, notamment lorsqu'elle concerne un projet réel [Neuthaus&Vizedom, 2013]. Avec la numérisation des documents et ensuite l'accessibilité à un nombre important de données et de textes sur le web, l'engouement pour la construction semi-automatique d'ontologie à partir de textes, est devenu important. Plusieurs techniques et outils de disciplines différentes, telles que les techniques de traitement automatique des langues, d'apprentissage automatique ou de raisonnement logique ont été proposés [Buitelaar *et al.* 2005], [Navigli *et al.* 2011]. Les propositions qui en résultent sont prometteuses mais ne sont pas exploitables en l'état [Gruninger&Obrst, 2014].

Dans ce contexte, nous nous sommes intéressés à la construction semi-automatique des ontologies à partir de textes. Nous avons tout d'abord proposé un cadre pour la comparaison des approches et outils de construction semi-automatique en réalisant une extension des travaux existants [Park *et al.* 2011]. *Nous avons ensuite utilisé ce cadre pour étudier certaines méthodes et outils existants. En outre, en l'absence d'un outil gratuit et libre, qui englobe la majorité des étapes du processus de construction pour la langue française, nous avons adapté Text2onto, outil pour la construction semi-automatique à partir des textes en anglais, à la langue française.*

Nous avons participé, dans le cadre de trois projets (lot de travail UEML du Rex INTEROP, ISTA3 et KIFANLO), au développement d'ontologies. Nous avons suivi un processus de construction descendante (dans UEML), ascendante (dans ISTA3) ou mixte (dans KIFANLO), en suivant l'approche générale MethOntology et une méthode basée sur une ontologie noyau pour une conceptualisation manuelle ou semi-automatique d'une ontologie formelle (UEML) ou semi-formelle (ISTA3, KIFANLO). Dans ces projets, nous avons mis en évidence l'intérêt et le rôle important qu'une ontologie noyau formelle peut jouer dans un processus de construction et nous en avons dégagé des meilleures pratiques qui ont orienté nos travaux vers le développement d'une approche semi-automatique de construction intégrée à la validation.

La validation d'ontologie est une étape incontournable dans le processus de construction d'ontologie. Cette étape devient de plus en plus complexe lorsque la taille des ontologies devient de plus en plus grande.

Dans nos travaux, nous considérons la conceptualisation comme un processus composé de deux sous-processus menés en parallèle et en coopération : (1) le processus d'extraction et (2) le processus de validation. Nous considérons que le processus de validation doit être réalisé le plus

tôt possible, sur des sous-ensembles des résultats de l'extraction, afin d'éviter la propagation des problèmes dans l'ontologie et rendre difficiles et complexes leur identification et correction. En outre, nous considérons que le processus de validation est composé (1) d'un sous-processus qui évalue la qualité d'une ontologie, et (2) d'un deuxième processus pour améliorer cette qualité, si elle n'est pas suffisamment bonne, en identifiant et en enlevant ses causes.

Plusieurs méthodes et moyens ont été proposés pour l'évaluation de la qualité d'une ontologie. Citons par exemple, l'évaluation qualitative par des experts, l'évaluation qualitative selon les résultats d'une application utilisant cette ontologie, et l'utilisation des anti-patrons correspondant à des défauts avérés ou potentiels [Harzallah *et al.* 2015]. Enfin, des outils ont été développés pour identifier certains problèmes logiques dans une ontologie et proposer des solutions, comme les raisonneurs logiques. Mais plusieurs autres problèmes ne sont pas encore traités.

Nous avons abordé le problème d'évaluation d'ontologie suivant deux approches. Dans la première approche, nous avons considéré le concept de « problème » comme une facette de l'évaluation de la qualité d'une ontologie. Nous avons montré la variabilité des types de problèmes proposés et le manque de consensus sur leur définition. *Nous avons proposé une nouvelle typologie semi-formelle des problèmes (la majorité de ses classes est formalisée en logique de description) qui standardise les types de problèmes existants et les classes selon deux axes : les erreurs vs les situations indésirables et les problèmes sociaux vs les problèmes logiques.* Cette typologie est inspirée des travaux sur la qualité des modèles conceptuels (*i.e.* le modèle SEQUAL [Krogstie *et al.* 1995]). Elle couvre les problèmes cités dans la littérature mais elle est aussi généralisable à de nouvelles classes.

En nous basant sur cette typologie et les liens entre ses problèmes, nous avons proposé un ordre pour l'identification de de certains problèmes dans une ontologie permettant d'éviter leur propagation dans cette une ontologie et d'optimiser le processus de sa validation.

Pour l'identification des problèmes proprement dit, la majorité des travaux s'est intéressée à l'identification et à la correction des problèmes d'insatiabilité en faisant intervenir ou non un acteur social. En revanche, peu de travaux ont porté sur la détection des problèmes sociaux et particulièrement sur les problèmes de contradiction sociale.

Nous nous sommes intéressés à ce type de problème et nous avons proposé des anti-patrons partiels et une heuristique basés sur des liens formels entre ces problèmes et les problèmes d'insatisfiabilité afin de les identifier tout en minimisant l'intervention humaine.

Nous nous sommes également intéressés au problème d'« Ontologie Plate », notamment lorsque cette ontologie est construite d'une façon automatique. Nous avons proposé une méthode pour classifier des concepts extraits sous les concepts noyaux, afin d'aider à corriger ce problème.

Notre deuxième approche pour la validation d'ontologie s'intègre dans le cadre de l'évaluation d'une ontologie par l'utilisation. Nous avons proposé une approche de construction et de validation d'ontologie quand cette dernière est construite pour l'annotation d'objets à l'aide d'un méta-modèle. Cette approche est basée sur une liste de règles formelles générée à partir d'un méta-modèle d'annotation et appliquée pendant le processus d'annotation avec cette ontologie.

Mesures sémantiques de comparaison d'objets. Dans le troisième axe, nous nous sommes intéressés à la définition et à l'analyse des mesures sémantiques appliquées à une ontologie pour améliorer l'exploitation de cette dernière dans certaines applications. L'évolution rapide et perpétuelle de l'environnement des organisations et l'hétérogénéité de ses données dans des bases de documents ou dans le web ont rendu difficile la recherche d'information, la comparaison d'objets (gènes, documents, cartes géographique…), l'interopérabilité des systèmes...

[Zargayouna *et al.* 2016]. L'utilisation des ontologies est apparue comme une solution incontournable pour expliciter la sémantique de ces données et de ces vocabulaires dans ces applications. En effet, une ou plusieurs ontologies pourraient être utilisées pour expliciter la sémantique des connaissances des systèmes (nous parlerons dans la suite d'une annotation des données, connaissances ou objets avec les artefacts d'une ontologie). Ensuite des mesures sémantiques pourraient évaluer la proximité sémantique de ces données, connaissances ou objets et aider ainsi à identifier des correspondances sémantiques entre eux [Harzallah&Berio, 2015]. Cependant, le choix d'une mesure sémantique adéquate à une ontologie et à l'objectif de l'application qui l'utilise n'est pas une tâche facile.

Plusieurs mesures sémantiques pour la comparaison de deux concepts d'une même ontologie ont été proposées dans la littérature. *Nous sommes les premiers à avoir proposé un cadre pour comparer et analyser ces mesures et en définir si nécessaire une nouvelle [Blanchard et al. 2008]. Nous avons étendu ce cadre aux mesures sémantiques comparant des ensembles de concepts et ensuite aux mesures sémantiques comparant des graphes sémantiques d'une même ontologie. Nous avons ainsi obtenu un cadre unifiant pour la définition et l'analyse des mesures sémantiques de comparaison d'objets annotés par une ontologie. Un objet pourrait être une donnée, un produit, une carte géographique, un gène, une image... et il pourrait être annoté par un concept unique, un ensemble de concepts ou un graphe sémantique d'une ontologie.*

Une nouvelle famille de mesures sémantiques pour la comparaison des graphes sémantiques a été proposée après une étude des travaux en fouille et en comparaison des graphes. Dans cette nouvelle famille de mesures, nous avons défini le graphe sémantique commun maximum à deux graphes sémantiques et le contenu informationnel d'un graphe sémantique.

Nous avons appliqué notre cadre dans des domaines différents : (1) dans le domaine de l'ingénierie des modèles pour comparer des constructs des langages où ces derniers ont été annotés par des ensembles de concepts et ensuite par des graphes sémantiques; *(2) dans le domaine des métiers d'entreprise pour l'interopérabilité de systèmes hétérogènes où les données d'un système ont été annotées par des concepts uniques, et (3) sur des réseaux sociaux pour comparer des profils d'utilisateurs et détecter des communautés où ces profils ont été annotés par un concept ou un ensemble de concepts d'une ontologie.*

La Fig. 1 illustre nos contributions majeures et les liens entre elles.

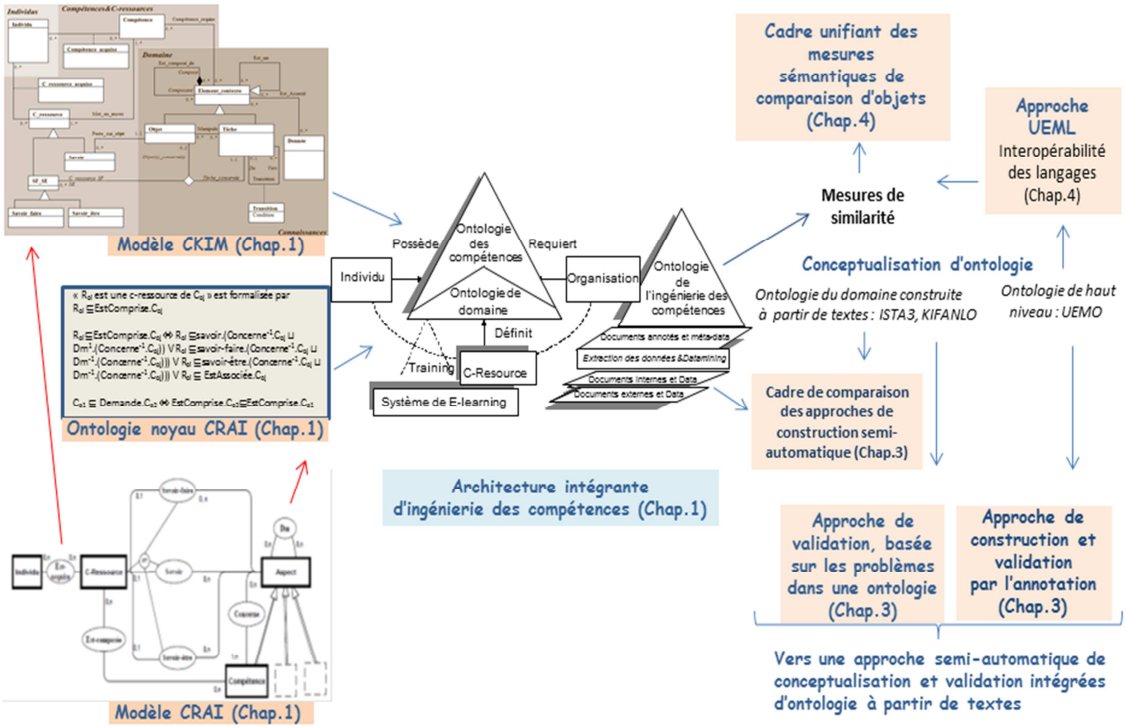

Fig. 1 – Nos contributions majeures et les liens entre elles

Dans ce mémoire d'HDR, nous détaillons nos différentes contributions en les positionnant par rapport à des travaux connexes de l'état de l'art. Nous concluons sur la pertinence de nos travaux et nous présentons notre projet de recherche et des perspectives liées au challenge du web des données.

Dans le chapitre 1, intitulé **« De l'ingénierie des compétences vers l'ingénierie des connaissances »,** nous exposons les liens entre données, connaissances et compétences dans l'ère de la masse de données et dans la nouvelle discipline des sciences des données (section 1.2). Nous présentons ensuite différentes méthodes de modélisation des connaissances ainsi que nos contributions en modélisation des compétences. Dans la section 1.3, nous présentons des techniques d'ingénierie des connaissances provenant de différentes disciplines et nous discutons de leur intérêt pour l'ingénierie des compétences avant de présenter notre architecture intégrante pour l'ingénierie des compétences. En conclusion, nous évoquons la possibilité d'étendre notre architecture pour qu'elle intègre l'ingénierie des compétences, connaissances et données.

Dans le chapitre 2, intitulé **« De la construction manuelle d'ontologie vers la construction semi-automatique »**, nous présentons dans la section 2.2 des approches de construction d'ontologie. Ensuite, nous nous intéressons particulièrement à l'étape de conceptualisation. Nous présentons des méthodes et approches existantes pour accomplir cette tâche et leurs caractéristiques. Nous discutons de leur insuffisance pour accomplir cette tâche dans un cas réel (section 2.3). Nous considérons ensuite des méthodes, techniques et outils de conceptualisation semi-automatique d'ontologie et nous proposons un cadre pour les comparer (section 2.4). Nous partageons dans la section 2.5 notre expérience en construction d'ontologie dans trois projets (lot

de travail UEML du Rex INTEROP, ISTA3 et KIFANLO) en mettant en exergue le rôle clé d'une ontologie dans ces projets (section 2.5).

Dans le **chapitre 3**, intitulé « **Vers une approche semi-automatique de construction et validation intégrées d'ontologie »**, nous présentons nos contributions pour le développement de deux approches de validation d'ontologie : une première approche de validation par l'identification des problèmes pouvant nuire à la qualité d'une ontologie et une deuxième approche de validation par l'utilisation d'une ontologie dans une application. Pour la première approche, nous proposons une nouvelle typologie évolutive des problèmes pouvant nuire à la qualité d'une ontologie (section 3.3.1.1) ; un ordre d'identification des problèmes, qui optimise le processus de validation d'ontologie (section 3.3.1.2) ; des anti-patrons partiels pour l'aide à l'identification du problème de « Contradiction sociale » (section 3.3.1.3) ; et une méthode pour l'aide à la correction du problème d'« Ontologie Plate » utilisant des règles inductives définies à partir d'une ontologie noyau (section 3.3.1.4). La deuxième approche est une approche pour la construction et la validation d'une ontologie développée pour l'annotation (3.3.2). En conclusion, nous articulons nos contributions pour aller vers une approche semi-automatique de construction et validation intégrées d'ontologie, basée sur une ontologie noyau formelle.

Dans le chapitre 4, intitulé **« Cadre unifiant des mesures sémantique de comparaison d'objets »,** nous illustrons la problématique de comparaison d'objets définis avec des données hétérogènes dans le cadre de l'approche UEML[1] (section 4.2). Nous présentons ensuite les trois familles de mesures sémantiques de comparaison que nous avons définies, tout en les positionnant par rapport aux travaux existants (sections 4.3, 4.5 et 4.6). Nous détaillons en particulier la troisième famille de mesures qui permet de comparer deux objets annotés chacun par un graphe sémantique. Dans la section 4.4, nous présentons notre approche pour l'approximation du contenu informationnel, élément principal et commun à la définition de ces trois familles de mesures. Nous présentons dans la section 4.8 la variabilité des résultats des mesures sémantiques ayant chacune une approximation différente du contenu informationnel. Nous traitons également la variabilité des résultats d'une mesure donnée avec l'évolution de l'ontologie à laquelle elle est appliquée. Nous présentons dans la section 4.9 notre cadre pour la définition des mesures sémantiques et ses paramètres avant de clore ce chapitre avec des perspectives d'amélioration de nos contributions.

Dans le Chapitre 5, intitulé **« Conclusion, Projet de recherche et Perspectives »,** nous montrons la pertinence de nos contributions en ingénierie des connaissances pour le challenge du web des données. Nous présentons en deuxième partie notre projet de recherche et quatre perspectives liées à nos travaux dont deux portent sur la construction et la validation semi-automatiques d'ontologie, une porte sur notre cadre unifiant des mesures sémantiques et la dernière fait le lien entre nos travaux et l'interopérabilité des logiciels de gestion et de pilotage de production dans le département QLIO de l'IUT de Nantes.

[1] Unified Enterprise Modelling Language (http://www.uemlwiki.org/)

Chapitre 1 : De l'ingénierie des compétences vers l'ingénierie des connaissances

Sommaire

1.1 Introduction

La connaissance est une notion à la fois utilisée dans le langage courant et étudiée dans plusieurs disciplines (philosophie, psychologie, sciences cognitives, sociologie, informatique, *etc.*). Elle est considérée depuis un certain temps comme un capital immatériel permettant d'améliorer le bon fonctionnement d'un individu ou d'une organisation. Les processus de son acquisition, de sa modélisation, de son exploitation et de son enrichissement sont devenus impératifs et omniprésents dans la vie de l'être humain ou d'une organisation. L'ingénierie des connaissances est une discipline créée pour bien mener ces différents processus. La gestion des connaissances (ou knowledge management) fait partie de l'ingénierie des connaissances et porte sur les processus d'exploitation des connaissances en orientant, organisant, et coordonnant les activités et les processus destinés à amplifier l'utilisation et la création des connaissances.

L'ingénierie des connaissances a longtemps porté sur les connaissances des méthodes et tâches d'une organisation afin d'améliorer leur réalisation. La diversité et la multiplication rapide des connaissances ont amené les organisations à les extraire, les modéliser, les capitaliser et les stocker afin de créer leur « mémoire[2] ». Des systèmes de bases des connaissances ont été développés pour accéder vite et à moindre coût aux connaissances requises pour accomplir des tâches, répondre aux besoins des clients, *etc.* Avec la numérisation des documents et ensuite la prolifération des données sur le web, l'objet de l'ingénierie des connaissances s'est étendu aux connaissances sur autrui, sur un contexte, sur un environnement, sur des communautés… Les connaissances sont de plus en plus extraites à partir des données, des textes, des traces, *etc.* dans des bases de données, dans des applications ou dans le web. Une nouvelle discipline a émergé : il s'agit de la science des données qui porte sur l'étude des techniques pour traiter les données (et de masses des données) pour l'ingénierie des différents types de connaissances.

Plusieurs types de connaissances peuvent être définis et utilisés pour améliorer le fonctionnement d'un système d'information, d'un système intelligent ou d'une organisation ou pour répondre aux besoins (réels ou latents) d'un individu, d'un groupe d'individus ou d'une société. Tout d'abord, nous pouvons distinguer les connaissances descriptives des connaissances prédictives. Les connaissances descriptives sont (1) des connaissances du domaine, (2) des connaissances procédurales comme des connaissances de méthodes ou des connaissances d'inférences, et (3) des connaissances factuelles. Les connaissances prédictives portent sur l'identification d'un comportement, d'un besoin ou des relations possibles ou futures entre entités (objets, individus, actions, fonctions, organisations, sociétés…).

Nous pouvons distinguer aussi les connaissances élicitées des connaissances tacites. Les connaissances élicitées sont extraites à partir des compétences des experts, en observant ces experts effectuer des tâches ou en explicitant avec eux leurs connaissances tacites. Elles explicitent la façon dont les experts accomplissent une tâche ou atteignent un objectif. Ce que les experts connaissent, savent faire ou sont capables de faire sont des connaissances portant sur leurs compétences. Les connaissances sur les compétences peuvent être extraites aussi à partir des textes ou des traces sur le web (*e.g.* CVs, rapports de projets) ou des réseaux sociaux (*e.g.* LinkedIn). Les compétences elles-mêmes peuvent être acquises « sur le terrain » par des formations ou en appliquant des connaissances.

L'intérêt pour la gestion des compétences a augmenté avec les mouvements importants de départ à la retraite. Il s'est amplifié avec l'utilisation croissante du web pour diffuser des CVs (support des compétences) et la possibilité d'un traitement semi-automatique de ces derniers.

[2] Mémoire d'organisation

Cependant, peu de modèles opérationnels étaient disponibles pour la représentation des compétences et pouvant être utilisés pour leur extraction et leur gestion.

Depuis nos travaux de thèse de doctorat, nous nous sommes intéressés à la modélisation des compétences des ressources humaines, comme un aspect clé d'une organisation à prendre en compte pour améliorer sa performance. Nous avons proposé un modèle formel pour la modélisation des compétences : CRAI (Competency Resource Aspect Individual Model).

En arrivant dans l'équipe DUKe, nous avons étudié les compétences comme un type de connaissances afin d'y appliquer des techniques de l'ingénierie des connaissances. Nous avons débuté ce travail avec la définition des liens entre compétences et connaissances et la proposition du modèle CKIM (Competency and Knowledge Integrated Model) pour l'articulation de ces deux concepts. Nous avons ensuite transformé le modèle CRAI en une ontologie noyau des compétences (i.e. ontologie noyau CRAI) car la représentation conceptuelle du modèle CRAI n'est pas suffisante pour raisonner sur les compétences et inférer des nouvelles connaissances. En deuxième lieu, nous nous sommes intéressés aux techniques et outils de l'ingénierie des connaissances pour l'ingénierie des compétences. Nous avons défini une architecture intégrant la réalisation de différents processus d'ingénierie des compétences. Cette architecture, basée sur une ontologie de domaine et notre ontologie noyau CRAI des compétences, inventorie les sources des connaissances et les techniques de leur ingénierie pour extraire, modéliser et exploiter les compétences.

Dans la première partie de ce chapitre, nous rappelons les trois notions de compétence, connaissance et donnée et nous exposons des liens entre celles-ci (sections 1.2.1 et 1.2.2). Nous présentons ensuite des méthodes et modèles pour représenter différents types de connaissances et nous montrons l'intérêt des ontologies pour la modélisation des connaissances (section 1.2.3). Nous exposons ensuite nos contributions en modélisation des connaissances : le modèle CRAI, l'ontologie noyau CRAI et le modèle CKIM (section 1.2.4).

Dans la deuxième partie de ce chapitre, nous présentons des techniques en ingénierie des connaissances pouvant être appliquées en ingénierie des compétences (section 1.3.1). Certaines de ces techniques ont été utilisées dans d'autres travaux que nous avons réalisés et qui sont exposés dans les chapitres suivants. Nous présentons ensuite notre architecture pour l'ingénierie des compétences (section 1.3.2). Nous clôturons ce chapitre par une perspective d'extension de notre architecture à une architecture intégrante pour l'ingénierie des compétences, connaissances et données.

1.2 Vers une modélisation intégrée des compétences, connaissances et données

1.2.1 Données et connaissances

Les données sont au cœur de l'activité des êtres vivants, des objets du monde, des associations humaines... Elles peuvent porter sur des domaines formels (mathématiques), des domaines du vivant, des domaines financiers... Elles peuvent se trouver dans les livres, dans des bases des données, dans des modèles, dans le web… Elles peuvent se présenter sous la forme d'un terme ou d'une valeur numérique dans un texte d'un livre ou sur le web, dans un programme d'une application, dans les cellules d'un tableau, dans des modèles, dans les images, *etc.* Les données sont partout autour de nous : elles décrivent ce qu'on dit, ce qu'on pense, les résultats d'une machine, d'un programme, d'un questionnaire… Elles peuvent être brutes, semi-structurées ou structurées : des données non structurées ou semi structurées dans des textes, des images, des

sons, *etc.* ou des données structurées suivant un modèle, un tableau, une matrice, *etc.* Elles peuvent être contextualisées ou non comme des termes ou des valeurs numériques isolées.

Plusieurs travaux ont étudié les liens entre les notions de donnée et connaissance et ont montré leur proximité [Ganascia, 1998], [Schreiber *et al.* 2000], [Abiteboul, 2012]. Abiteboul [2012] considère que ces notions sont très proches. Il les définit comme suit : « *Une donnée est une description élémentaire, typiquement numérique pour nous, d'une réalité. C'est par exemple une observation ou une mesure. À partir de données collectées, de l'information est obtenue en organisant ces données, en les structurant pour en dégager du sens. En comprenant le sens de l'information, nous aboutissons à des connaissances, c'est-à-dire à des faits considérés comme vrais dans l'univers d'un locuteur, et à des « lois » (des règles logiques) de cet univers*».

Plusieurs domaines de recherche se sont intéressés à structurer des données, à les gérer ou à en extraire des connaissances : développement des bases de données, traitement automatique des langues, apprentissage automatique, fouille des données… Actuellement, tous les types de données peuvent être produits, numérisés, stockés, traités et échangés par une application, une organisation ou une personne. Le challenge actuel est de pouvoir exploiter les grandes masses de données. La disponibilité de cette masse de données de provenances différentes, obtenue en temps réel, est une source pertinente pour prévoir les besoins de chacun, et s'y adapter. Plusieurs domaines d'application ont profité de cette masse de données pour extraire et exploiter des connaissances : la recommandation (ou la personnalisation en fonction des affinités et des proximités entre clients et produits) ; l'évaluation de la réputation (ou de l'expertise où on évalue la qualité des informations sur le web) ; la notation de produits par les internautes ; la collaboration entre internautes pour réaliser collectivement une tâche qui les dépasse individuellement ; le crowdsourcing qui met des humains au service de systèmes informatiques ou la personnalisation… [Stanford University, 2016], [Abiteboul *et al.* 2017].

Une nouvelle discipline a émergé : la Science des Données. La définition de cette discipline n'est pas encore bien établie, elle est encore en cours. Elle peut être considérée comme la capacité ou la généralisation de l'extraction de connaissances à partir des données et l'utilisation de ces connaissances pour prendre des décisions. Elle emploie des techniques et des théories provenant de diverses disciplines : les mathématiques, les statistiques, la théorie de l'information, la fouille de données, l'apprentissage automatique, la reconnaissance de formes, la visualisation, la modélisation d'incertitude, le stockage de données, la compression de données, le calcul à haute performance, etc. [Schut&O'Neil, 2013].

1.2.2 Connaissances et compétences

Dans nos travaux de thèse, nous avons relié la notion de compétence à celle de connaissance [Harzallah, 2000]. La compétence est définie par la mise en œuvre combinée de savoir, savoir-faire ou savoir-être dans un contexte pour atteindre un objectif [Le Boterf, 97], [Harzallah, 2000]. La compétence est acquise en partie à partir des connaissances mais aussi avec de l'expérience et sur le terrain. En plus, les connaissances peuvent être extraites ou acquises à partir des compétences [Nonaka&Takauchi, 1997]. En effet, en observant comment une personne réalise une tâche, on peut formaliser ses connaissances procédurales pour la réalisation de cette tâche . Par ailleurs la description des compétences, par exemple ce que sait faire un individu ou une organisation ou ce qu'il est nécessaire de faire pour atteindre un objectif, est aussi de la connaissance.

Avec un exemple issu du domaine de la production manufacturière, nous illustrons les liens entre les notions de compétence, connaissance et donnée. Dans un atelier de fabrication de chaises, la quantité mensuelle fabriquée par chaque machine est une donnée (1600 chaises rouges

par la machine M1 et 2000 chaises bleues par la machine M2), le temps d'arrêt mensuel pour panne de chaque machine est une donnée (4 heures d'arrêt pour M1 et 10 minutes d'arrêt pour M2) et le temps mensuel consacré à la maintenance préventive de chaque machine est aussi une donnée (le temps consacré à la maintenance préventive de M1 est égale à 0 et celui à la maintenance préventive de M2 est égale à 1 heure). La règle suivante est de la connaissance : « Si on réalise la maintenance préventive d'une machine, on augmente probablement sa productivité ». Enfin, la compétence « Etre compétent dans la maintenance préventive des machines d'une entreprise donnée » implique qu'on connaisse les procédures pour faire de la maintenance préventive (une connaissance) dans cette entreprise, qu'on a déjà appliqué ces procédures (savoir-faire), et qu'on sait les adapter aux différentes machines de cette entreprise et dans des situations différentes (savoir-être).

En théorie, les liens entre les notions de connaissance et compétence sont plus au moins bien établis. En pratique, les notions de compétence et connaissance sont parfois confondues (*i.e.* elles sont parfois utilisées comme des synonymes) ou traitées séparément que ce soit en entreprise ou en recherche. En effet, la gestion des connaissances est en général réalisée dans un service (par exemple, le service R&D), et celle des compétences dans un second (par exemple, le service R.H.). Ces deux services fonctionnent indépendamment et communiquent peu entre eux. Au niveau de la recherche, lors de nos travaux sur cette problématique, les connaissances et les compétences commençaient à être évoquées ensemble [Dieng *et al.* 1998], [Probst *et al.* 2000], [Von Krogh, 2000], [Mille, 2001]. Cependant, peu de travaux de recherche ont été réalisés pour étudier leur complémentarité bien qu'elles soient fortement liées.

Nous avons mis en exergue dans cette section et dans la section précédente que les connaissances peuvent être extraites à partir des données ou des compétences et que les compétences peuvent être acquises à partir des connaissances ou extraites à partir des données. Pour cela, nous pensons que les techniques d'ingénierie et la modélisation sont deux moyens pertinents pour représenter et formaliser les liens entre ces trois notions.

1.2.3 Modélisation des connaissances

Depuis plusieurs années, les définitions de la notion de connaissance ont convergé vers la définition suivante : la connaissance est l'information qui prend une certaine signification dans un contexte donné et permet d'effectuer des tâches, l'information étant une ou plusieurs données structurées et contextualisées [Ermine, 2000] [Mille, 2001] [Collin, 2001]. La modélisation des connaissances est reconnue depuis longtemps, comme un moyen pertinent pour aider à extraire les connaissances, les expliciter, améliorer leur exploitation et raisonner avec ou sur elles. Plusieurs méthodes ont été proposées. Nous présentons dans la suite, certaines d'entre elles développées pour des objectifs différents et des domaines variés.

1.2.3.1 Représentation des connaissances dans un système à base de connaissances

A partir des années 80-90, on s'est intéressé à définir des modèles génériques pour aider les experts à expliciter leurs connaissances et capitaliser ainsi les connaissances d'entreprise. Ces modèles peuvent ensuite être utilisés dans des systèmes à base des connaissances (SBC). Un SBC est un système informatique qui utilise des connaissances et des procédures d'inférence à propos d'un domaine afin d'arriver à résoudre des problèmes dont la résolution nécessite une grande expertise humaine [Gonzalez&Dankel, 1993]. Des méthodes et des modèles ont été définis pour aider à acquérir, formaliser et gérer ces connaissances (*i.e.* expertises humaines). Les méthodes les plus connues sont KADS (Knowledge and Analysis Design Support) ou CommonKADS [Wielinga *et al.* 1992], [Schreiber *et al.* 2000] et MKSM (Method for Knowledge System

Management) ou MASK [Ermine *et al.* 1996]. D'autres méthodes sont un peu moins répandues, notamment REX (Retour d'EXpérience), KOD, SPIRIT, SAGACE, SPIRAL, AKM, GAMeth (Global Analysis Methodology) [Saad *et al.* 2002].

Dans ces différentes méthodes, trois types de connaissances sont généralement distingués : des connaissances du domaine, des connaissances factuelles et des connaissances de tâches (ou méthodes) [Wielinga *et al.* 1992], [Dieng *et al.* 1998]. Les connaissances du domaine représentent des connaissances peu évolutives qui décrivent les concepts d'un domaine et les relations entre eux. Les connaissances des tâches décrivent les tâches qui s'effectuent dans un système ou un domaine et le processus de leur accomplissement avec les conditions d'entrée, et de sortie et les contrôles. Ces connaissances sont souvent appelées des connaissances de résolution de problèmes. Les connaissances factuelles sont des faits qui ont eu lieu à un certain moment exprimant un changement d'état d'un ou plusieurs concepts du domaine.

La méthode CommonKADS développée pour concevoir des systèmes à base des connaissances (SBC) s'organise autour de six modèles (i.e. Modèle de l'organisation, Modèle de tâches, Modèle d'agents, Modèle de communications, Modèle de conception et le Modèle de connaissances) et une démarche pour les définir [Schreiber *et al*. 2000]. Les trois premiers modèles (Modèle de l'organisation, Modèle de tâches et Modèle d'agents) permettent de représenter et analyser l'environnement où le SBC va être utilisé ainsi que les tâches qui vont être accomplies à l'aide de ce SBC et les agents qui vont les réaliser. Le modèle de communications représente les transactions entre les agents impliqués dans le modèle des tâches. Le modèle de conception permet la spécification technique du SBC.

Le modèle de connaissances est une spécification des connaissances et des données requises pour l'accomplissement des tâches. Ce modèle distingue trois types de connaissances : (1) les connaissances du domaine (*i.e.* connaissances propres au domaine sur lequel va porter le SBC), (2) les connaissances de tâches qui décrivent les buts du SBC et les tâches et leur décomposition en sous-tâches pour atteindre ces buts, (3) les connaissances d'inférence décrivant les étapes d'inférences (*i.e.* méthode de résolution de problème) utilisant des connaissances du domaine pour accomplir chaque sous-tâche (cf. Fig. 1.1).

Quant à la méthode MASK (ou MKSM) c'est une méthode d'analyse préalable à la mise en place d'un système opérationnel de gestion de connaissances [Aries *et al.* 2008]. MASK procède par recueil des connaissances auprès des sources de connaissances de l'entreprise (*e.g.* experts, spécialistes) ou à partir de documents de références. Elle formalise ensuite ces connaissances à l'aide d'un ensemble de modèles et le résultat obtenu est un livre des connaissances. Elle propose sept modèles. Un modèle général qui délimite le domaine dont on souhaite capitaliser ses connaissances et qui définit sa finalité. Deux modèles sont utilisés pour aborder les activités d'un domaine : les modèles dits « de phénomène » et « d'activité ». Ils correspondent à l'analyse du système du point de vue contextuel : De quoi parle une connaissance ? Dans quelle activité est-elle mise en œuvre ? Quels sont les phénomènes que cherche à maîtriser une activité ?

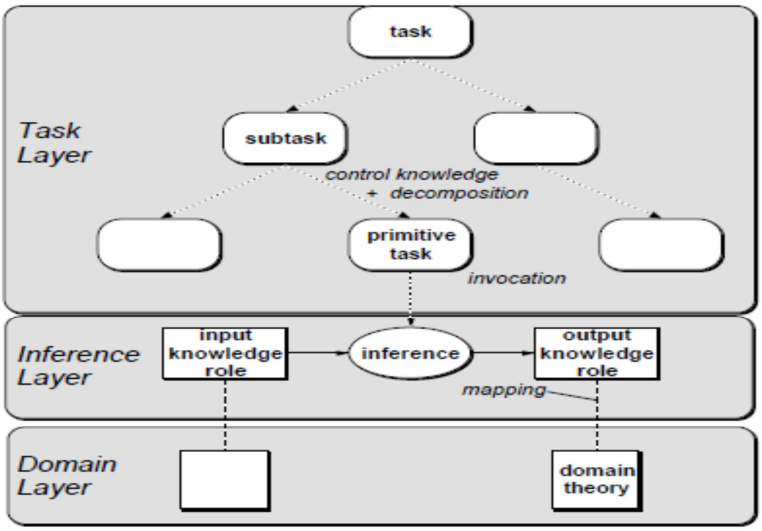

Fig. 1.1 - Modèle des connaissances [Fensel *et al.* 1998]

Dans MASK, deux modèles sont utilisés pour décrire les connaissances. Le premier est le modèle de tâches. Il décrit les aspects dynamiques et le raisonnement d'un domaine. Le deuxième est le modèle de concepts représentant les connaissances statistiques d'un domaine. Il structure et classifie les concepts et les objets d'un domaine en utilisant des relations de spécialisation et des relations de composition entre eux et en définissant leurs attributs et leurs instances. Deux modèles sont utilisés pour décrire l'histoire de la construction de la connaissance. Le modèle de l'historique représente les relations que la connaissance entretient avec d'autres sous-systèmes. Le modèle de lignée donne une image des évolutions des objets et concepts du système.

1.2.3.2 Représentation des connaissances en modélisation d'entreprise

Dans le domaine de la modélisation d'entreprise, des modèles ou méta-modèles sont proposés pour représenter les différents aspects de l'entreprise et les utiliser comme support pour analyser l'entreprise et la réorganiser ou pour développer des outils pour informatiser son système d'information ou des outils de simulation de son fonctionnement. Nous pouvons citer les modèles de GRAI, IDEFs, CIMOSA, GERAM, [Vernadat, 1996] et BPMN [OMG, 2011]. Par exemple, dans CIMOSA (CIM Open System Architecture) on préconise de modéliser l'entreprise suivant quatre vues complémentaires (*i.e.* les vues Fonction, Information, Ressource et Organisation) et on offre des primitives et des patrons pour modéliser chacune de ces vues.

Les modèles d'entreprise se focalisent plus sur les connaissances déclaratives, pouvant être utilisées pour analyser et simuler les processus métiers de l'entreprise et aider à les optimiser (cf. Fig. 1.2). Par rapport aux méthodes MASK ou CommonKADS, les approches de modélisation d'entreprise ne proposent pas de diagramme de concepts (comme dans MASK), de modèle des connaissances du domaine ni de modèle des connaissances de tâche (comme dans CommonKADS). Ces deux derniers ressemblent à une ontologie du domaine. Enfin, les approches de modélisation d'entreprise ne proposent pas non plus de modèle des connaissances d'inférence comme dans CommonKADS.

Par ailleurs, des ontologies portant sur le domaine de l'entreprise ont été développées, comme EnterpriseOntology [Uschold *et al.* 1996] ou TOVE [Fox *et al.* 1998].

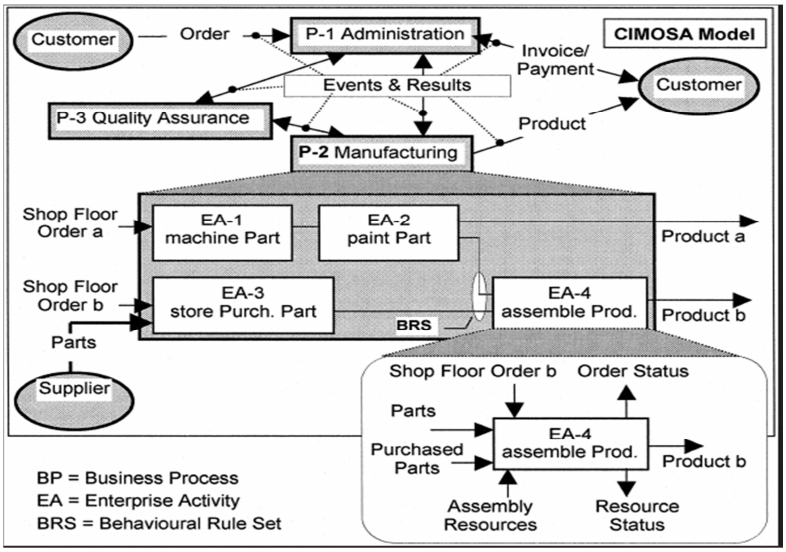

Fig. 1.2 – Modélisation du processus « gestion d'une commande » avec CIMOSA [Vernadat, 1999]

1.2.3.3 Représentation des connaissances causales ou probabilistes

Parmi les modèles les plus connus et les plus pertinents pour la représentation des connaissances incertaines, nous pouvons citer les cartes cognitives et les réseaux bayésiens. Une carte cognitive est une manière de représenter des assertions causales ou d'une façon générale des relations d'influences sur un domaine. Elle se représente sous la forme d'un graphe orienté, étiqueté et acyclique où les nœuds représentent les éléments d'un domaine et les arcs représentent les liens d'influence entre eux (cf. Fig. 1.3) [Chauvin, 2010]. Un réseau bayésien est un modèle graphique probabiliste pouvant représenter des relations probabilisées entre les variables de ses nœuds. Il permet de représenter des dépendances incertaines entre connaissances [Pearl, 1988], [Leray, 2006].

Dans une carte cognitive ou dans un réseau bayésien, la sémantique des labels associés aux nœuds n'est pas définie. De plus, un nœud n'est pas nécessairement un concept d'un domaine.

Ces deux types de modèles pourraient devenir inexploitables, s'il y a des ambiguïtés dans la sémantique de leurs nœuds, de leurs arcs ou de leurs variables.

1.2.3.4 Représentation des connaissances et Ontologies

Pour les différents types de modèles présentés ci-dessus, on a identifié un besoin d'expliquer la sémantique des concepts, entités, classes, nœuds, relations, arcs ou variables utilisés dedans, parce qu'un label associé à un de ces éléments peut être compris de plusieurs façons (*i.e.* un polysème) et que plusieurs de ces éléments peuvent avoir des labels différents tout en étant quand-même identiques (*i.e.* synonymes). Les problèmes de polysémie et de synonymie peuvent nuire à l'exploitation de ces modèles pour la définition des connaissances.

Les ontologies ont été introduites en ingénierie des connaissances, principalement, pour régler ces problèmes : (1) pouvoir représenter les connaissances et expliciter leur sémantique d'une façon qu'elles soient compréhensibles de la même façon par des humains et des machines, (2) pouvoir réutiliser des modèles des connaissances, (3) et pouvoir raisonner sur et avec ces connaissances. Les ontologies peuvent être utilisées seules ou associées à d'autres représentations des connaissances afin de définir la sémantique des composants de ces dernières. Par exemple, on

peut utiliser une ontologie pour définir la sémantique des différents nœuds d'une carte cognitive [Chauvin, 2010].

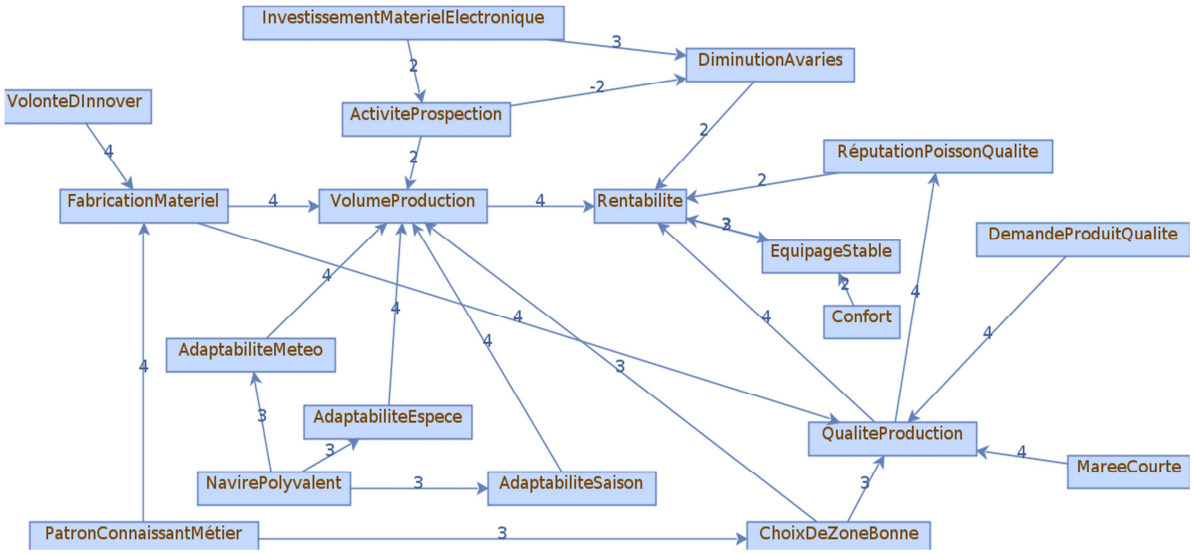

Fig. 1.3 - Carte cognitive construite à partir d'un entretien avec un pêcheur dans le projet KIFANLO

En ingénierie des connaissances, la définition informelle d'une ontologie la plus connue et la plus largement acceptée est celle proposée par Gruber [1993] : « Une spécification formelle et explicite d'une conceptualisation partagée ». Une « conceptualisation » est une vue abstraite et simplifiée du monde que l'on souhaite représenter dans un but donné. Le qualificatif « partagée » associé à la notion de conceptualisation indique qu'elle est le résultat d'un compromis qui satisfait une communauté de personnes. La spécification « formelle » d'une conceptualisation est la formalisation de celle-ci dans un langage pouvant être interprété par les machines. La spécification « explicite » indique que dans une ontologie ses différents composants ou artefacts et les contraintes sur leurs utilisations sont explicitement définis.

Du point de vue sémiotique, une ontologie est considérée comme un objet sémiotique utilisé par des machines sur lequel et avec lequel on peut raisonner. Elle doit donc être formelle. Elle est considérée aussi comme un objet interprété, utilisé et transformé par l'humain, qui doit pouvoir la comprendre et l'exploiter [Aimé&Charlet, 2014]. Le composant principal d'une ontologie est le concept et d'un point de vue sémiotique, ce composant peut être défini : par des notions, par des propriétés ou par des instances (*i.e.* triangle sémiotique [Richards&Ogden, 1989]). *Les instances* d'un concept sont l'ensemble des êtres, objets ou faits auxquels le concept fait référence : il s'agit d'une définition extensionnelle d'un concept. Cependant, l'extension d'un concept peut être vide, notamment dans le cas des concepts abstraits comme, par exemple le concept de « Sérénité ». Par ailleurs, la frontière entre les notions de concept et d'instance n'est pas toujours facile à définir. Elle dépend parfois du degré de granularité avec lequel on veut conceptualiser un domaine. Les propriétés d'un concept sont les caractéristiques essentielles et communes aux êtres, objets ou faits auxquels le concept fait référence et qui lui permettent de se différencier des autres concepts : il s'agit d'une définition intentionnelle d'un concept. Enfin, un concept peut être associé à un ou plusieurs termes ou étiquettes (labels) synonymes de son label : il s'agit d'une définition d'un concept par des notions. Parfois, la distinction entre un terme et un concept n'est

pas prise en compte dans une ontologie : plusieurs termes représentant le même sens se trouvant dans une même ontologie. Dans ce cas, il ne s'agit pas vraiment d'une ontologie mais plutôt d'une terminologie.

Une ontologie peut comprendre d'autres artéfacts : les relations, les attributs et les axiomes. Les relations dans une ontologie décrivent des associations entre deux concepts. Comme pour un concept, une relation peut être aussi définie par des notions, par des propriétés ou par des objets. La relation taxonomique ou de subsomption, notée dans la suite *est-un* est un type particulier de relation entre concepts. C'est une relation transitive, asymétrique et réflexive. Elle définit un ordre partiel entre les concepts d'une ontologie et permet leur organisation dans une hiérarchie avec la règle d'héritage des propriétés des pères par les fils et celle d'héritage des objets des fils par les pères.

Les attributs d'un concept définissent des propriétés intrinsèques de ce concept acquises par tous ses objets. Un ou plusieurs termes synonymes peuvent être associés à un attribut.

Les axiomes dans une ontologie sont des contraintes sur des concepts ou des relations afin de compléter leur définition et de mieux formaliser leur sens. En plus des axiomes de subsumption entre concepts, d'autres types d'axiomes peuvent être définis à l'aide des opérateurs d'équivalence ou de négation, des quantifications universelle et existentielle, *etc.*

Dans la littérature, plusieurs définitions formelles de la notion d'ontologie ont été proposées. Elles dépendent principalement des éléments qui permettent de définir ses concepts et ses relations (i.e. objets, propriétés ou notions), de la méthode de sa construction et de l'objectif de son utilisation. Nous pouvons distinguer trois définitions. La première définition a pour origine la définition informelle d'une ontologie selon Gruber [1993] et basée sur la formalisation de la notion de conceptualisation [Guarino *et al.* 2009]. La deuxième définition considère une ontologie comme une structure et précise ses composants et les relations entre eux [Stumme, 2003]. La troisième définition élargit la première à une ontologie lexicalisée (*i.e.* un ou plusieurs labels sont associés à chaque composant de l'ontologie), à une ontologie peuplée (i.e. des objets sont associées à chaque composant de l'ontologie) ou à un système axiomatique.

Les ontologies sont classées généralement en trois catégories selon le degré d'abstraction de leurs concepts : les ontologies fondationnelles (« upper ontology » ou « top ontology), les ontologies de domaine et les ontologies d'application [Guarino, 1997]. Une ontologie fondationnelle comprend des concepts abstraits (*e.g.* temps, espace,) censés permettre de structurer, par spécialisation, la conceptualisation de n'importe quel domaine [Declerck *et al.* 2012]. Une ontologie de domaine comprend les concepts d'un domaine particulier (*e.g.* le domaine médical). Une ontologie d'application comprend les concepts d'un domaine d'une application et les concepts de tâches de ce domaine.

La catégorie d'ontologies noyaux (core ontology en anglais) a été rajoutée comme une catégorie intermédiaire entre les ontologies fondationnelles et les ontologies de domaine [Gangemi&Borgo, 2004]. Une ontologie noyau comprend des concepts génériques et commun à certains domaines. Elle pourrait être construite à partir d'une ontologie fondationnelle. Burita *et al.* [2012] considèrent qu'une ontologie noyau d'un domaine(s) est composée du minimum de concepts et de relations nécessaires pour décrire/comprendre d'autres concepts de ce domaine(s) (« In philosophy, a core ontology is a basic and minimal ontology consisting only of the minimal concepts required to understand the other concepts » [Burita *et al.* 2012]). L'objectif prinicpal d'une ontologie noyau est d'aider à construire ou à aligner des ontologies de domaine [Gangemi&Borgo, 2004 ; Guarino, 2009 ; Sherp *et al.* 2009]. Pour cela, Gangemi *et al.* [2004] et Sherp *et al.* [2009] considèrent qu'une ontologie noyau doit être formelle et comprendre le maximum d'axiomes.

Dans nos travaux, nous avons adopté la définition d'une ontologie noyau selon Burita et al. Nous supposons que chaque concept ou relation non taxonomique d'une ontologie d'un domaine construite à partir d'une ontologie noyau est une spécialisation respectivement d'un concept ou d'une relation de cette ontologie noyau. D'autres relations entre des concepts subsumés par un même concept noyau peuvent être rajoutées.

1.2.4 Notre contribution en modélisation des connaissances

Les compétences sont un type de connaissances qui porte sur ce qu'on sait, ce qu'on peut faire ou ce qu'on est capable de faire pour atteindre un objectif dans un contexte donné. L'ingénierie des compétences est un domaine plus récent que celui de l'ingénierie des connaissances. Des nomenclatures (ou référentiels) listant des compétences, par exemple le ROME et le Cigref, ou des bases de compétences spécifiques à des entreprises ont été développées. Cependant, avant nos travaux, il n'y avait ni modèle exploitable définissant le concept de compétence d'une façon unifiée et formelle, ni modèle le liant au concept de connaissance. Nous nous sommes donc intéressés au concept de compétence depuis nos travaux de thèse et nous avons proposé le modèle des compétences CRAI. En se basant sur ce modèle, nous avons défini une ontologie noyau des compétences comme support pour définir une ontologie des compétences sur laquelle et avec laquelle on peut raisonner. Nous avons également proposé le modèle CKIM pour représenter les connaissances et les compétences d'une façon intégrée. Dans la suite nous présentons ces contributions.

1.2.4.1 Modèle CRAI pour la modélisation des compétences

Pendant nos travaux de thèse, nous avons étudié des travaux de sociologues, gestionnaires et cogniticiens portant sur le concept de compétence (*e.g.* [Levy-leboyer, 1996], [Le Boterf,1997] [Lucia&Lepsinger, 1999], [Pfeffer&Sutton, 2000]) et nous avons adopté la définition suivante pour modéliser ce concept : « une compétence est la combinaison et la mise en œuvre de savoirs, de savoir-faire et de savoir-être dans un contexte pour accomplir une mission ou une tâche ». Les savoirs, savoir-faire et savoir-être sont considérées comme des ressources de la compétence. Par exemple, la compétence C1 « Etre compétent en Java » nécessite la combinaison et la mise en œuvre d'un ensemble de c-ressources dont nous pouvons citer :

- S1 « Connaître les principes des langages objets »,
- S2 « Connaître l'API Java »,
- S3 « Connaître la syntaxe de Java »,
- S4 « Connaître les principes d'un algorithme »,
- SF1 « Savoir installer Java »,
- SF2 « Savoir gérer les exceptions en Java »,
- SF3 « Savoir programmer en Java»,
- SE1 « Savoir appréhender une difficulté ».

Nous avons développé un modèle générique des compétences, nommé le modèle CRAI (Competency, Resource, Aspect, Individual) défini avec le formalisme entité-relation étendu (cf. Fig. 1.4) et ensuite formalisé en théorie des ensembles [Harzallah, 2000], [Harzallah&Vernadat, 2002], [Harzallah *et al.* 2002], [Harzallah&Berio, 2004], [Harzallah *et al*. 2006].

Le modèle CRAI est composé de quatre entités principales : « Compétence », « C-ressource » (Savoir, Savoir-faire et Savoir-être), « Aspect » et « Individu ».

L'entité Aspect permet de définir l'objectif, la mission, la tâche… ou l'objet sur lequel porte une compétence et son contexte. Elle représente le domaine pour lequel on veut définir ses compétences. Afin de bien structurer ce domaine, l'entité Aspect peut être spécialisée ou décomposée en d'autres entités et des relations entre ces entités peuvent être rajoutées. Fig. 1.5 représente un exemple illustratif de spécialisation de l'entité Aspect quand ce dernier représente le domaine de l'informatique.

L'entité « Compétence » comprend les compétences requises par une organisation ayant le domaine considéré ou acquises par les individus de l'entité « Individu ». L'entité « C-ressource » comprend les ressources des compétences. Nous avons nommé les ressources d'une compétence « c-ressources » pour les distinguer des ressources financières ou humaines d'une organisation.

Le modèle CRAI comprend six relations principales. La relation « concerne » est définie entre une compétence et un aspect du domaine (par exemple, C1« être compétent en Java » concerne l'aspect Java).

La relation « Savoir » relie une c-ressource à un aspect du domaine (e.g. S3 « Connaître la syntaxe de Java » est reliée à l'aspect « Syntaxe_de_Java » par « Savoir »). La relation « Savoir-faire » relie une c-ressource à un aspect du domaine avec une propriété de type verbe (*e.g.* SF1 « Savoir installer Java » est reliée à « Java » par « Savoir-faire » et le verbe « installer »). La relation « Savoir-être » relie une c-ressource à un aspect du domaine avec une propriété de type verbe (*e.g.* SE1 « Savoir appréhender une difficulté » est reliée à « Problème » par « Savoir-être » et le verbe « appréhender »). Ces trois relations permettent de déterminer pour une compétence donnée, à l'aide d'une requête simple, certaines c-ressources nécessaires pour la mise en œuvre de cette compétence. En effet, il s'agit d'une requête qui identifie les c-ressources reliées par une des trois relations « Savoir », « Savoir-faire » et « Savoir-être » à l'aspect qui est lui-même relié à cette compétence par la relation « Concerne ». Par exemple, ce type de requête identifie pour la compétence C1 « être compétent en Java » les deux c-ressources SF1 « Savoir installer Java » et SF3 « Savoir programmer en Java». Cependant, elle ne permet pas d'identifier toutes les c-ressources de C1.

D'autres c-ressources sont nécessaires pour acquérir une compétence donnée mais ne portent pas nécessairement sur l'aspect de cette compétence mais plutôt sur un aspect lié à son aspect. Par exemple, S1 « Connaître les principes des langages objets » est nécessaire pour acquérir C1, mais elle porte sur l'aspect « principes des langages objets ». Ce dernier est nécessaire à maitriser pour maitriser Java. Pour identifier ce type de c-ressources, nous avons rajouté la relation Dm (Demande la maîtrise) entre les éléments de Aspect et nous avons rajouté la règle suivante :

> Si (une compétence porte sur « aspect1 », une c-ressource porte sur « aspect2 », et « aspect1 » Dm « aspect2 »), alors cette c-ressource est une c-ressource de cette compétence.

Enfin, d'autres c-ressources sont nécessaires à la mise en œuvre d'une compétence et ne peuvent pas être identifiées à l'aide de la requête ou la règle présentées ci-dessus (*e.g.* S4 et SE1 pour C1). Pour relier ces c-ressources à une compétence, nous avons rajouté la relation « Est-associée » entre une compétence et une c-ressource.

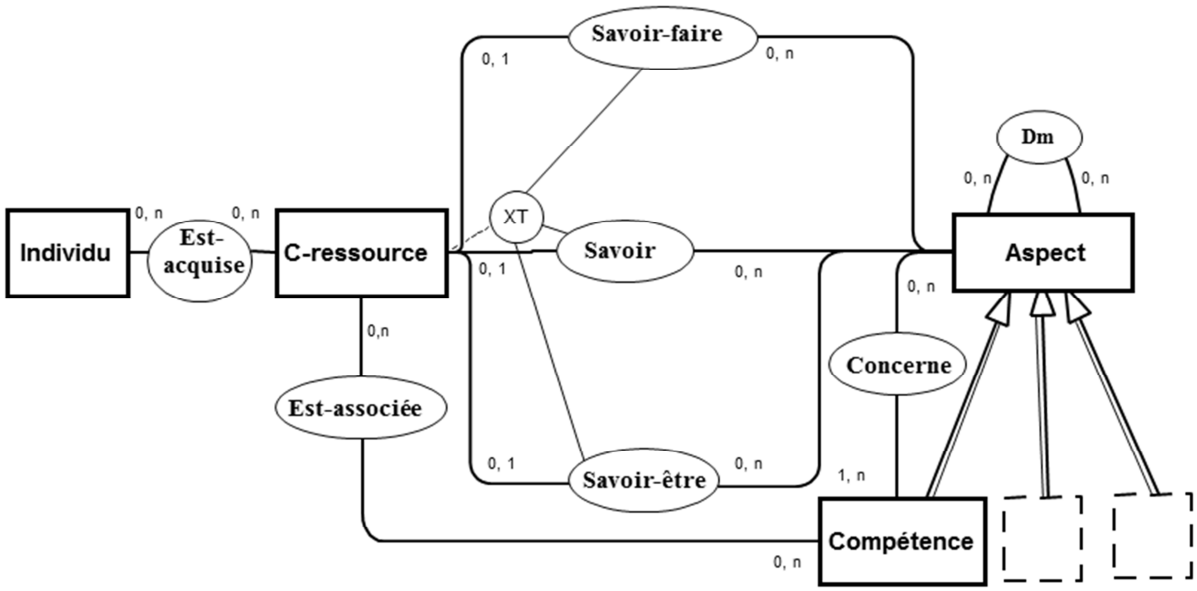

Fig. 1.4 - CRAI-Modèle entité relation étendu des compétences [Harzallah, 2000]

En résumé, les relations « Savoir », « Savoir-faire » et « Savoir-être » entre une c-ressource et un aspect, la relation Dm entre les aspects, et la relation « Est-associée » entre une compétence et une c-ressource permettent de définir et de déterminer toutes les c-ressources d'une compétence. La relation « Est-acquise » permet de déterminer les individus ayant acquis une compétence donnée, en vérifiant si ils ont acquis ses c-ressources nécessaires.

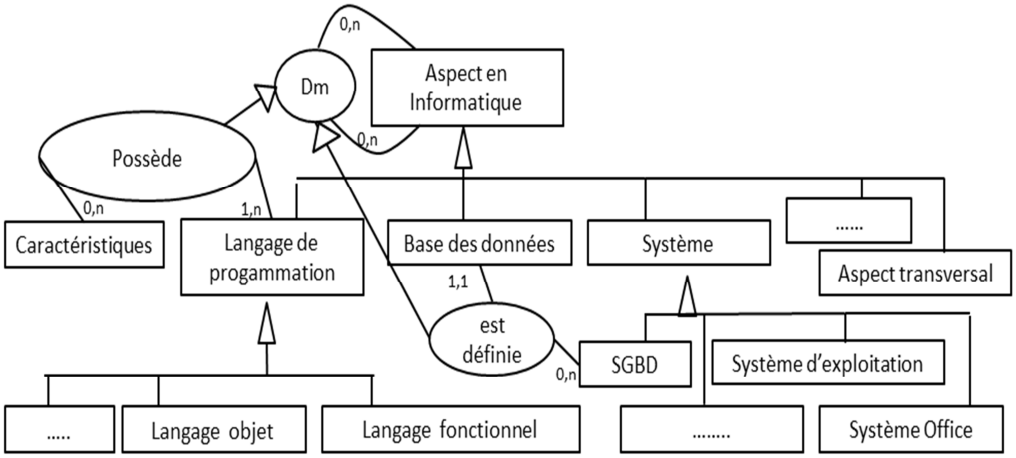

Fig. 1.5 - Exemple illustratif de la spécialisation de l'aspect informatique

Le modèle CRAI est considéré parmi les premiers modèles génériques et semi-formels des compétences. Il offre une représentation formelle et standardisée des compétences individuelles permettant une gestion unifiée des compétences acquises ou requises. En effet, outre que les liens entre compétences, c-ressources et aspects sont bien formalisés, une description standard des compétences et c-ressources peut être générée à partir du modèle CRAI à l'aide de patrons que nous avons définis. Par exemple, le patron « Savoir + Verbe +Aspect » est utilisé pour générer la

description d'une c-ressource de type savoir-faire. Par exemple, il permet de générer la description de la c-ressource SF1 qui est reliée à « Java » par « Savoir-faire » et la propriété du type verbe « installer » : « Savoir installer Java » [Harzallah *et al.* 2006].

Le Modèle CRAI est accompagné d'une liste de directives pour le déployer dans une organisation spécifique pour concevoir son système d'information de compétences et le maintenir. Il prend en compte et aide à formaliser les quatre processus d'ingénierie des compétences dans une organisation : (1) processus d'identification des compétences requises, (2) processus d'acquisition des compétences ; (3) processus d'identification et d'évaluation des compétences acquises ; et (4) processus d'exploitation des compétences [Harzallah *et al.* 2006].

Nous avons déployé ce modèle sur un exemple académique dans [Harzallah *et al.* 2006] en montrant comment l'utiliser pour définir des compétences et des c-ressources requises ou acquises pour le domaine de la maintenance et comment utiliser cette définition pour mettre en place les quatre processus de gestion des compétences. Lors de nos travaux de thèse, nous avons développé une application pour la gestion des compétences du service de maintenance de l'entreprise Tremery, basée sur ce modèle [Harzallah&Vernadat, 2002].

Dans un de nos derniers articles sur les compétences [Berio *et al.* 2011], nous avons comparé le modèle CRAI à plusieurs autres propositions pour la modélisation des compétences, comme par exemple celles dans les travaux de Marreli [1998], Lucia&Lepsinger [1999], Laukkanen&Helin [2005], De Coi *et al.* [2007], Sitthisak *et al.* [2007] ou de Sampson&Fytros [2008] ou les importantes initiatives de HR-XML[3] ou de IMS RDCEO[4] (IMS Reusable Definition of Educational Objective). Dans ces approches, les notions de compétence et de c-ressource ne sont pas distinguées. En plus, la relation entre une compétence et un aspect n'est pas défini. Ces lacunes en formalisation ne permettent pas, par exemple, de savoir ce qu'il faut apprendre ou la formation nécessaire pour acquérir une compétence ; ou de faire des rapprochements sémantiques entre compétences. Par ailleurs, Sampson & Fytros, (2008) ont indiqué le manqué d'un consensus sur la définition du concept de compétence et ont proposé, une définition proche de la nôtre !

Enfin, le modèle CRAI est toujours cité aux niveaux national et international comme une référence pour développer des systèmes pour la gestion des compétences [Rauffet *et al.* 2014], [Elia&Margherita, 2015], [Guerrero *et al.* 2015].

1.2.4.2 CKIM pour la modélisation intégrée des compétences et connaissances

Dans les années 2000, peu des travaux ont cherché à modéliser ou à gérer les concepts de compétence et de connaissance d'une façon intégrée, alors qu'il y a des liens évidents entre eux. Nous nous sommes intéressés tout d'abord à définir ces liens et clarifier la frontière entre ces deux concepts. Ensuite, nous avons étudié les modèles associés à ces deux concepts et établi le modèle CKIM (Competency and Knowledge Integrated Model) : modèle de gestion intégrée des compétences et connaissances. CKIM défini en UML, comprend trois parties principales [Vergnaud, 2003], [Vergnaud *et al.* 2004] (cf. Fig. 1.6). U*ne première partie* (Compétences&C-ressources) modélise les compétences et c-ressources d'un domaine (elle fait partie du modèle CRAI). *Une deuxième partie* (Domaine) modélise les connaissances d'un domaine sous la forme d'objets qui pourraient être transformées, utilisées… par des tâches comme dans CommonKADS. Il peut s'agir des tâches des processus métiers d'une organisation ou des tâches d'un SBC. Dans CKIM, les connaissances d'inférence sont représentées en partie. Seulement l'ordre de réalisation des tâches et les conditions de transition d'une tâche à une autre sont définis grâce à la classe

[3] Consortium HR-XML, HR-XML, versions 2.5 to 3.1, 2010. (www.hr-xml.org).
[4] IMS, RDCEO, *http://www.imslobal.org/competencies*, 2010.

« Transition ». A chaque concept du domaine, une donnée de type document, image, vidéo, *etc.* pourrait être associée permettant de compléter sa définition, l'enrichir ou représenter la ressource à partir de laquelle il a été extrait. Enfin, *une dernière partie* (Individus) modélise les individus dont on évalue les compétences acquises.

La construction de ce modèle débute par la définition du lien entre connaissances et compétences, à savoir le domaine sur lequel portent les compétences et les connaissances.

En plus de la définition des connaissances d'un domaine, CKIM permet de modéliser et d'expliciter les c-ressources et compétences. Par exemple :

- l'instanciation de la relation « *Porte_sur_objet* » entre la c-ressource « $S1_{sous\_catégorie\_S=Savoir\text{-}théorique}$ » et l'objet « *Principes des langages objets* » définit le savoir théorique : Connaître les principes des langages objets ;
- l'instanciation de la relation entre les trois classes SF-SE, Objet et Tâche avec « $SF1_{sous\_catégorie\_SF = Savoir\text{-}faire\text{-}empirique}$ », « *Java*», et « Installer » permet de définir le savoir-faire : Savoir Installer Java.

Dans le master de N. Vergnaud [2003], nous avons implémenté le modèle CKIM en développant un prototype de gestion des compétences connecté (interfacé) à l'application ATHANOR[5] de gestion des connaissances en diagnostique. Ce prototype a permis d'étendre ATHANOR à une application de gestion intégrée des connaissances et compétences.

Le modèle CKIM rend explicite le lien entre une compétence et une connaissance en les associant à un même concept de Domaine. Il établit également à l'aide des concepts de Domaine les liens entre les compétences et les données ou ceux entre les connaissances et les données. En effet, les données (textes, documents, modèles...) sont associées aux concepts de Domaine sur lesquels portent les connaissances et les compétences. Enfin, dans ce modèle, le lien entre Donnée et Domaine est défini par la relation « Est_Associé ». Cette relation peut-être spécialisée en « Est_extrait de » ou « Est_représenté_par ».

[5] ATHANOR a été développé et commercialisé par la société PerformanSe (www.performanse.fr)

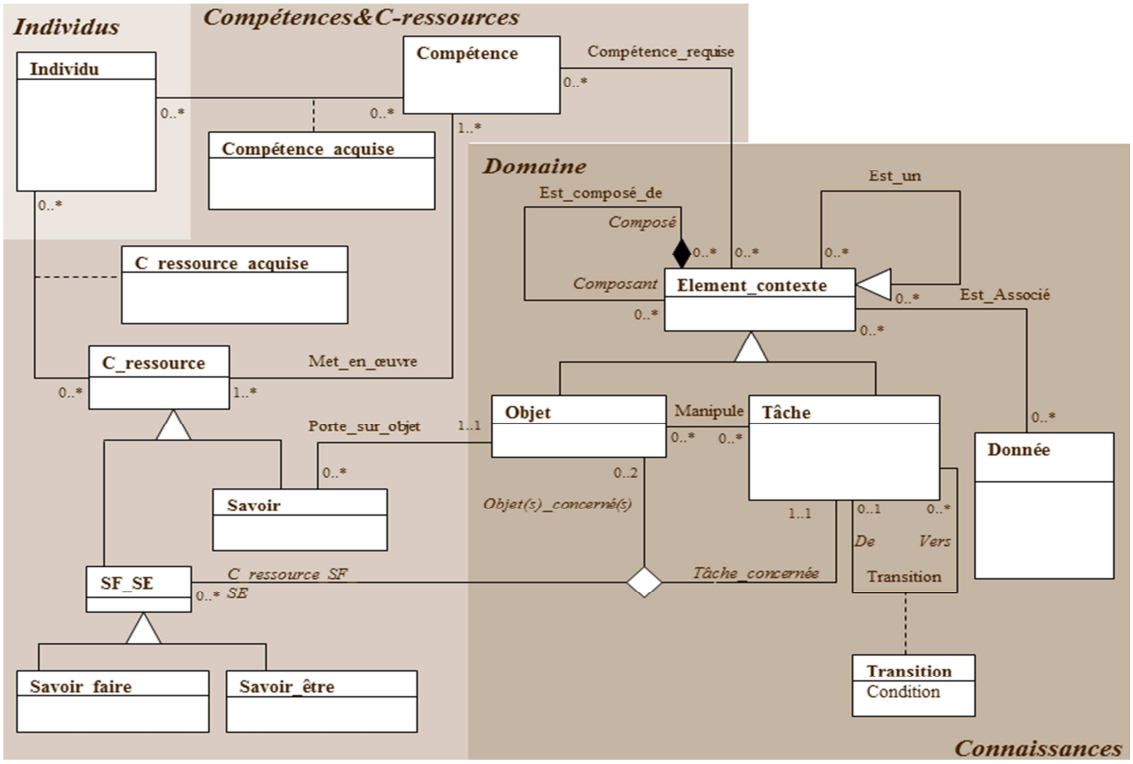

Fig. 1.6 - CKIM : Competency and Knowledge Integrated Model ([Vergnaud *et al.* 2004] avec modifications)

1.2.4.3 Ontologie noyau CRAI

Lors de la considération des connaissances métiers (*e.g.* mémoire d'entreprise), les ontologies sont avérées être une meilleure représentation des connaissances pour expliciter leur sémantique et raisonner sur elles. Dans ce contexte, nous nous sommes intéressés au développement d'ontologie des compétences en réutilisant le modèle CRAI.

Dans la littérature, on a souvent confondu une ontologie des compétences avec une ontologie du domaine des compétences : une compétence portant sur un concept d'un domaine (ou un aspect) est définie par le concept lui-même [Hiermann&Hôfferer, 2003], [Laukkanen&Helin, 2005], [Colucci *et al.* 2003a], [Corby *et al.* 2004]. Cette modélisation ne permet pas d'expliciter par exemple, ce que veut dire qu'un individu a acquis une compétence donnée.

Nous avons proposé une ontologie noyau des compétences en adaptant le modèle CRAI. Dans nos travaux, nous considérons qu'une ontologie noyau d'un domaine est composée du minimum de concepts et relations (nommés concepts et relations noyaux) nécessaires pour décrire ce domaine (cf. section 1.2.4.3). Tout concept d'une ontologie de ce domaine construite à partir de cette ontologie noyau est une spécialisation d'un de ses concepts noyaux.

L'ontologie noyau CRAI est composée de trois concepts noyaux : Domaine (ce concept correspond à l'entité Aspect dans le modèle CRAI), Compétence et C-ressource. Le concept Domaine représente le domaine sur lequel portent les compétences de l'ontologie à définir. Il est spécialisé en concepts et des relations entre eux pour définir le module Domaine ou une ontologie du domaine [Posea&Harzallah, 2004], [Harzallah, 2004], [Berio&Harzallah, 2007].

Un concept subsumé de Compétence est associé à un ou à plusieurs concepts du module Domaine par la relation Concerne. Un concept subsumé de C-ressource est défini en l'associant à un seul concept subsumé de l'ontologie du domaine s'il s'agit d'un savoir (par la relation Savoir) et à un deuxième concept subsumé de Tâche, dans les deux autres cas (par les relations Savoir-faire ou Savoir-être). Tâche est un concept subsumé de Domaine et représente les tâches du domaine considéré. Certains concepts subsumés de C-ressources peuvent être reliés directement à un concept subsumé de Compétence, par la relation EstAssociée. Comme dans le modèle CRAI, nous avons rajouté la relation Dm entre les concepts du domaine pour représenter la relation de demande de maîtrise entre concepts.

Le rajout d'une relation de spécialisation ente les compétences n'a pas de sens. Nous avons donc défini une nouvelle relation hiérarchique entre elles : la relation Demande qui permet de structurer sémantiquement les compétences entre elles. Sa sémantique est la suivante :

Soit O une ontologie des compétences, C_{o1}, C_{o2} et C_{oj} trois concepts subsumés de Compétence et R_{oi} un concept subsumé de C-ressources.

C_{o1} Demande C_{o2} ça veut dire que l'acquisition de C_{o1} requiert l'acquisition de C_{o2}. Elle est formalisée comme suit :

$$C_{o1} \sqsubseteq \text{Demande}.C_{o2} \Leftrightarrow \text{EstComprise}.C_{o2} \sqsubseteq \text{EstComprise}.C_{o1} \qquad (1.1)$$

où

$$R_{oi} \sqsubseteq \text{EstComprise}.C_{oj} \Leftrightarrow R_{oi} \sqsubseteq \text{savoir}.(\text{Concerne}^{-1}.C_{oj} \sqcup \text{Dm}^{-1}.(\text{Concerne}^{-1}.C_{oj})) \text{ V } R_{oi} \sqsubseteq \text{savoir-faire}.(\text{Concerne}^{-1}.C_{oj} \sqcup \text{Dm}^{-1}.(\text{Concerne}^{-1}.C_{oj}))) \text{ V } R_{oi} \sqsubseteq \text{savoir-être}.(\text{Concerne}^{-1}.C_{oj} \sqcup \text{Dm}^{-1}.(\text{Concerne}^{-1}.C_{oj}))) \text{ V } R_{oi} \sqsubseteq \text{EstAssociée}.C_{oj} \qquad (1.2)$$

La relation Demande peut être rajoutée manuellement entre les compétences d'une ontologie ou inférée à partir des deux règles précédentes (1.1 et 1.2).

Par ailleurs, des similarités/rapprochements sémantiques entre compétences peuvent être déterminées à l'aide des mesures sémantiques. Pour les compétences d'une ontologie définie à partir de l'ontologie noyau CRAI, elles sont déterminées en calculant les similarités/rapprochements sémantiques entre les concepts sur lesquels portent les compétences.

Dans les travaux de master de V. Posea [2004], [Posea&Harzallah, 2004], nous avons adapté la distance sémantique de Sussna [1997] à la comparaison sémantique des compétences et nous l'avons appliquée à un exemple acadméique dans le domaine de l'informatique. Nous avons mené ensuite des travaux approfondis sur les mesures sémantiques de comparaison des concepts d'une ontologie (cf. Chap. 4).

1.3 Vers une architecture intégrante pour l'ingénierie des compétences, connaissances et données

La gestion des compétences avait souvent porté sur les compétences des individus à recruter sans s'intéresser à l'identification des compétences manquantes et requises dans une organisation, à l'évolution des compétences disponibles, à l'étude de l'adéquation entre les compétences acquises et celles requises par un projet, *etc.*

Nous avons cherché à identifier les différents processus d'ingénierie des compétences et à les gérer d'une façon intégrée et articulée. Tout d'abord, nous avons distingué quatre processus

d'ingénierie des compétences (chacun comprend un ou plusieurs sous-processus) : (1) le processus d'identification des compétences requises, à savoir l'identification et la définition des compétences requises par les tâches, les missions, les objectifs, *etc.* d'une organisation ; (2) le processus d'acquisition des compétences par les individus d'une organisation, par exemple, à l'aide de formations ; (3) le processus d'identification et d'évaluation des compétences acquises par des individus, compétences requises ou non par une organisation ; et (4) le processus d'exploitation des compétences déterminées dans les processus d'identification et d'évaluation des compétences, par exemple : Comment identifier les écarts entre les compétences acquises et requises ? Qui devrait suivre une formation donnée ? Comment identifier des individus clés (*i.e.* ayant des compétences clés) ?…

Ces quatre processus sont interconnectés. Les processus d'évaluation et d'acquisition des compétences peuvent être basés sur les compétences identifiées dans le premier processus. Des nouvelles compétences peuvent être identifiées suite aux processus d'évaluation ou d'acquisition des compétences. Enfin, l'exploitation des compétences est basée sur les connaissances portant sur les compétences acquises et requises. Des connexions dynamiques entre processus permettent de fiabiliser et améliorer l'efficacité d'un système de gestion des connaissances et par conséquent elles participent à l'amélioration de la performance de l'organisation qui utilise ce système. Toutefois, ces processus sont souvent gérés séparément, ce qui rend leur développement couteux et affaiblit leur efficacité. Le modèle CRAI, l'ontologie noyau CRAI et le modèle CKIM permettent de représenter les compétences et leurs c-ressources dans ces quatre processus d'une façon unifiée afin de les connecter sans effort.

Nous nous sommes intéressés à l'application des techniques de l'ingénierie des connaissances à l'ingénierie des compétences et à les répertorier dans une architecture d'ingénierie des compétences. Nos motivations étaient multiples. Tout d'abord, les compétences sont un type de connaissances pour lesquelles nous avons proposé une modélisation intégrée à celle des connaissances. Plusieurs techniques de l'ingénierie des connaissances peu connues dans le domaine de l'ingénierie des compétences pourraient être adaptées à l'ingénierie des compétences. En plus, les ressources à partir desquelles les compétences et les connaissances sont extraites sont similaires : experts, textes, web, traces, *etc.* Par exemple, les connaissances et les compétences peuvent être identifiées en suivant les traces d'un apprenant, d'un groupe d'individus ou d'une organisation, en analysant ses résultats, ses réponses, ses traces, ses projets, avec des techniques de traitement automatique de textes, d'apprentissage automatique…

Nous avons donc considéré les processus d'ingénierie des compétences et nous avons étudié les techniques d'ingénierie des connaissances qui peuvent s'appliquer à un ou à plusieurs de ces processus afin de les classifier dans une architecture intégrante d'ingénierie des compétences. Cette architecture est appelée intégrante pour deux raisons : elle intègre la réalisation des différentes processus d'ingénierie des compétences et elle répertorie les différentes techniques d'ingénierie connaissances et les ressources utilisées par ces techniques selon les processus et les sous-processus d'ingénierie des compétences qu'elles pourraient aider à accomplir.

1.3.1 Techniques pour l'ingénierie des connaissances

Il existe plusieurs travaux de recherche dans la littérature, ayant appliqué des techniques d'ingénierie des connaissances pour réaliser des processus d'ingénierie des compétences. Ces techniques peuvent être classées en trois catégories : (1) l'acquisition manuelle des connaissances, (2) l'extraction automatique des connaissances, et (3) les techniques de raisonnement.

L’acquisition manuelle des connaissances est centrée sur les experts du domaine et est aussi connue sous le nom « Knowledge Elicitation ». Cette méthode étant déjà évoquée dans la section 1.2.3, nous ne la considérons pas dans la suite. L’extraction automatique des connaissances est centrée plutôt sur l’utilisation d’algorithmes de découverte de la connaissance et est aussi connue sous le nom de fouille de données. Les techniques de raisonnement sont toutes les techniques qui s’appuient sur la notion d’épreuve, de propriété ou de théorème, qui sont associées à une logique quelconque et qui formalisent l’idée de déduction à partir d’un modèle ou d’une théorie donné(e). Ces trois catégories peuvent utiliser des techniques de modélisation des connaissances, des mesures sémantiques, de l’annotation… En outre, une technique peut utiliser une autre technique, par exemple certaines techniques d’extraction automatique demandent des techniques de raisonnement.

Dans la suite, nous introduisons ces techniques et nous indiquons dans la section 1.3.2 celles qui sont appliquées ou qui pourraient être appliquées à l’ingénierie des compétences.

1.3.1.1 Fouille de données

La fouille de données consiste à rechercher et à extraire des connaissances utiles et inconnues d’une grande quantité de données, en utilisant des méthodes automatiques ou semi-automatiques. Elle a initialement porté sur des données organisées sous la forme d’une matrice mettant en relation des individus et leurs attributs et est appelée dans ce cas fouille de données structurées. Elle s’est étendue ensuite à la fouille de textes, la fouille du web, la fouille d'images, la fouille audio, la fouille de flots de données, la fouille de séquences... La fouille de données structurées fait partie du domaine de recherche de l’extraction des connaissances dans des bases de données ou « Knowledge Dicovery in Databases ».

D’après Piatetsky-Shapiro [1991], l’extraction des connaissances à partir des données (ECD) désigne le processus non trivial conduisant à la découverte des informations implicites, inconnues et potentiellement utiles et compréhensibles à partir des données. Le processus de ECD comprend 5 étapes (cf. Fig. 1.7) : (1) la collecte de données, (2) leur nettoyage et leur prétraitement, (3) leur transformation et réduction, (4) l’application d’une méthode adéquate de fouille sur les données, et (5) l’interprétation et l’évaluation des connaissances extraites avant leur utilisation [Fayyad *et al.* 1996].

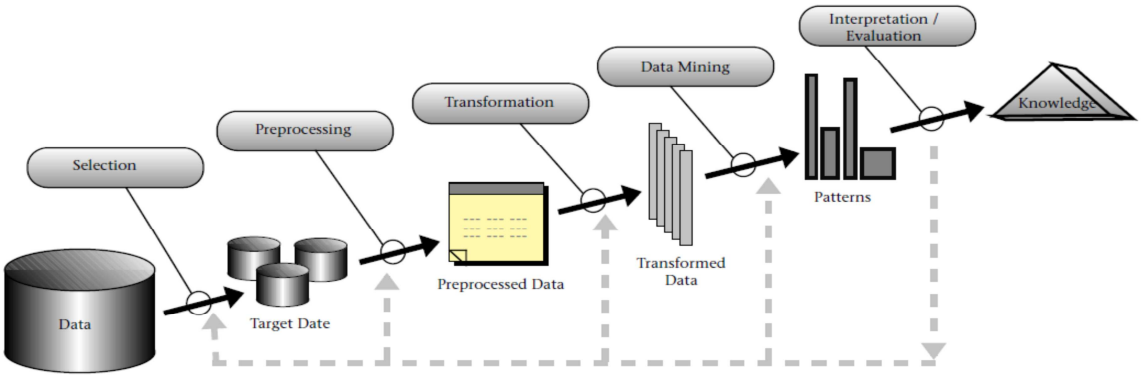

Fig. 1.7 - Etapes du processus de l’ECD [Fayyad *et al.* 1996]

Les méthodes de fouille de données sont traditionnellement classées en deux grandes familles : les méthodes descriptives, généralement non supervisées et les méthodes prédictives, généralement supervisées.

Les méthodes descriptives permettent d'organiser, de simplifier et d'aider à comprendre l'information sous-jacente d'un ensemble important de données [Gordon, 1981]. Elles ont pour objectif de proposer une structure inconnue des données. Généralement non supervisées, elles sont classées en trois groupes de méthodes [Ibekwe-SanJuan, 2007]. *Le premier groupe* correspond aux méthodes de classification automatique (i.e. clustering) et comprend les méthodes d'agrégation, les méthodes de partition et les cartes auto-organisantes de Kohonen. Les deux premières méthodes utilisent des mesures de similarités pour la classification des données. La dernière méthode se base sur les réseaux de neurones artificiels. *Le deuxième groupe* correspond aux règles d'association qui permettent de déterminer des relations entre les données. *Le dernier groupe* correspond aux méthodes factorielles comme l'analyse factorielle des correspondances ou l'analyse sémantique latente.

Les méthodes prédictives ont pour but de prédire ou d'expliquer le statut de nouveaux objets à partir de ceux des objets déjà connus. Elles vont s'intéresser par exemple à l'évaluation de la probabilité qu'un individu achète un produit plutôt qu'un autre, aux chances qu'un individu ayant visité une page d'un site web y revienne, etc. Généralement supervisées, elles se basent sur l'apprentissage automatique. L'apprentissage relève d'une démarche d'induction : on généralise les valeurs de classification à partir de l'observation d'un nombre limité d'exemples. Les méthodes d'apprentissage les plus utilisées sont : (1) la méthode des k-plus proches voisins, (2) la méthode à vecteurs supports, (3) les classifieurs bayésiens naïfs, (4) l'arbre de décision, et (5) les réseaux de neurones artificiels.

1.3.1.2 Fouille de textes

Nous nous intéressons particulièrement au processus d'extraction des connaissances à partir de textes à la fois comme un moyen d'extraire des compétences requises ou acquises à partir des textes (cf. section 1.3.2) et de construire semi-automatiquement des ontologies à partir de textes (cf. chapitres 2 et 3).

Le processus d'extraction des connaissances à partir de textes est décrit suivant quatre étapes (cf. Fig. 1.8) dans [Ibekwe-SanJuan, 2007], [Toussaint, 2011], [Nédellec, 2013]. La première étape est l'étape de prétraitement où les textes sont nettoyés et normalisés. La sélection des textes fait partie de cette étape. La deuxième étape est une étape de transformation qui consiste en l'occurrence à faire un choix des unités d'analyses (*e.g.* termes, phrases ou paragraphes) et de leurs caractéristiques. Cette étape s'effectue à l'aide des techniques de traitement linguistique de textes (*e.g.* on choisit des unités d'analyse selon leur catégorie morphosyntaxique, cf. la section suivante) ou/et des techniques de traitement statistique (*e.g.* on choisit un terme selon sa fréquence ou son tf-idf (Term frequency – Inverse Document Frequency) dans un contexte). Dans la troisième étape, il s'agit d'appliquer une ou plusieurs techniques de fouille sur ces unités. Les résultats de cette étape sont ensuite interprétés et évalués avant d'être exploités ou avant l'élaboration de connaissances. Il est évident que ce processus correspond bien au processus de l'ECD.

D'après Weiss *et al.* [2010], la fouille de textes ne serait qu'une dérivée de la fouille de données structurées. En effet, on peut introduire les unités d'analyse choisies dans une première dimension d'une matrice et des documents, des parties d'un document (*e.g.* section, phrase), des contextes, des caractéristiques, *etc.* dans la deuxième dimension de cette matrice. Une cellule de cette matrice représente la présence/absence ou la fréquence d'une unité d'analyse dans un

élément de la deuxième dimension. Cependant, la taille importante de ce type de matrice et le fait que cette dernière contient probablement beaucoup de zéros dans ses cellules rendent nécessaire l'adaptation d'une technique de fouille de données structurées avant de pouvoir l'appliquer à ce type de matrice.

Par exemple, l'analyse distributionnelle utilise ce type de modélisation matricielle pour déterminer des liens sémantiques entre des unités d'analyse. En effet, l'analyse distributionnelle se base sur la théorie de Harris selon laquelle les mots qui apparaissent dans les mêmes contextes tendent à avoir des significations proches. Deux types de contextes sont souvent considérés : les *contextes syntaxiques* (par exemple, argument d'un même verbe) *et les contextes graphiques* (par exemple, les mots dans le voisinage d'une unité d'analyse cible) [Périnet&Hamon, 2014]. L'analyse distributionnelle range les unités d'analyse dans la première dimension d'une matrice et leurs contextes dans sa deuxième dimension. Ensuite, l'application des méthodes de catégorisation ou d'agrégation à ce type de matrice permet de définir des classes sémantiques (*i.e.* unités d'analyse synonymes), des relations *est-un* entre des classes d'unités d'analyse ou d'autres types de liens/relations entre ces classes ou entre ses unités.

Des méthodes d'extraction des règles d'association permettent aussi de déterminer des relations sémantiques entre des unités d'analyse à partir de ce type de modélisation (*i.e.* la modélisation matricielle). Dans ce cas, les unités d'analyse de la première dimension d'une matrice sont considérées comme des objets (ou individus) et la deuxième dimension de cette matrice comprend des attributs de ces objets. Pour extraire des règles d'association à partir de ce type de matrice, on cherche tout d'abord à identifier des motifs fréquents. Un motif est composé d'un ou de plusieurs attributs et il est fréquent s'il appartient à plusieurs unités d'analyse. Les unités d'analyse ayant les mêmes motifs peuvent constituer des classes sémantiques. Ensuite, l'extraction des règles d'association se fait à partir des motifs fréquents. Une règle d'association est une implication de la forme M1 → M2 où M1 et M2 sont deux motifs et elle signifie que tout objet contenant M1 contient aussi M2 avec une probabilité égale à la confiance de cette règle. Pour filtrer des règles d'association, on peut utiliser des mesures de qualité ou des indices statistiques [Blanchard *et al.* 2004].

Enfin, des mesures de similarité sont appliquées à ce type des données (modélisation matricielle) pour déterminer des similarités sémantiques entre ces unités.

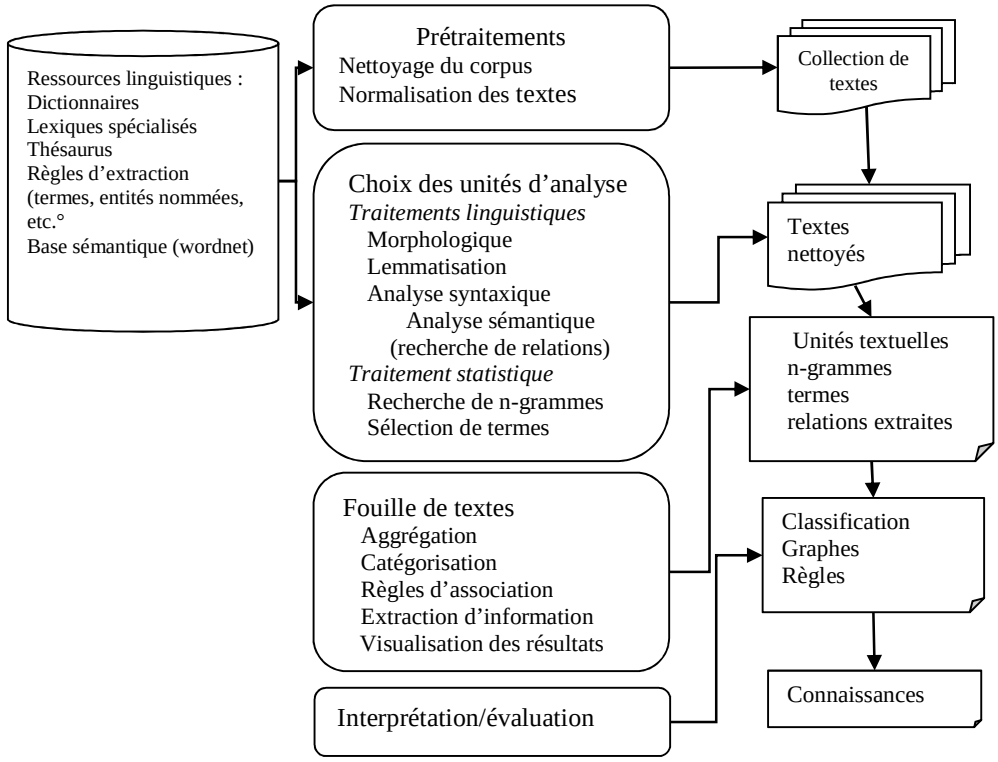

Fig. 1.8 - Processus d'extraction des connaissances à partir des textes ([Ibekwe-SanJuan, 2007] avec modification)

Le processus d'extraction des connaissances à partir de textes peut être adapté à des objectifs ou à des domaines particuliers en rajoutant ou en supprimant des étapes ou des techniques. Par exemple, le choix du label d'un concept associé à une classe sémantique est à rajouter à ce processus si on veut l'adapter à la conceptualisation d'ontologie (cf. chapitre 2).

1.3.1.3 Traitement automatique des langues

Le traitement automatique des langues (TAL) est une discipline à la frontière de la linguistique et de l'informatique. Le TAL porte sur l'analyse du contenu d'un corpus de textes afin de proposer un résumé, de faire la traduction automatique, de représenter le contenu d'une phrase ou d'un discours, d'extraire des connaissances sur un domaine, *etc.* Un processus de TAL s'effectue en 6 phases. La première phase est la segmentation d'un texte qui consiste à le découper en séquences ou en phrases. La deuxième phase est le découpage d'une phrase en mots (unité lexicale ou token), elle s'appelle généralement le « tokenization ». La troisième phase est la lemmatisation des unités lexicales qui consiste à associer à chacune sa forme lemmatisée : nom au singulier, verbe à l'infinitif, *etc.* La quatrième phase est l'analyse morphosyntaxique (tagging) qui consiste à attribuer à une unité lexicale une étiquette grammaticale (*e.g.* nom, verbe, adjectif) et un trait morphologique (*e.g.* genre, nombre). La cinquième phase est l'analyse des dépendances syntaxiques qui consiste à déterminer des relations de dépendance entre les unités lexicales. Diverses types de dépendances syntaxiques ont été proposées : des dépendances séquentielles qui ne sont pas vraiment syntaxiques [Bunescu&Mooney, 2006], des dépendances

entre unités lexicales en utilisant l'analyse syntaxique superficielle (i.e. shallow parsing) ou des dépendances identifiées à partir d'une représentation sous forme d'arbre den utilisant l'analyse syntaxique profonde [Culotta&Sorensen, 2004], [Moncecchi *et al.* 2010]. Cette phase permet de définir des unités d'analyse (e.g. des termes) et des relations entre elles. La sixième phase est l'analyse sémantique qui consiste à affecter une étiquette sémantique à chaque unité d'analyse et la reconnaissance des relations sémantiques entre elles.

Pour illustrer ces différentes phases, considérons la phrase « natural language processing tools extract terms automatically from texts ». Elle peut être découpée en 9 unités lexicales lemmatisées (phases 2 et 3) : « natural », « language », « processing », « tool », « extract », « term », « automatically », « from » et « text ». «tool », « language », « term » et « text » sont des noms, « Processing » est un gérondif, « extract » est un verbe, « automatically » est un adjectif et « from » est une préposition (Phase 4). « natural language processing tool » peut former une unité d'analyse et il est le sujet du verbe « extract », « term » est le complément d'objet direct du verbe « extract » et « text » est son complément d'objet indirect (Phase 5). « natural language processing tool » est un « tool » et « text » est une « textual ressource » (Phase 6). Fig. 1.9 représente un arbre syntaxique de cette phrase, l'étiquetage grammatical[6] est réalisé avec l'outil NLTK[7].

Plusieurs outils ont été proposés pour réaliser les 5 premières phases de ce processus. Des environnements complets d'analyse syntaxico-sémantique ont été développés : GATE, ANLT (Alvey Natural Language Tools), NLTK, les outils de Stanford NLP Group[8], *etc.*

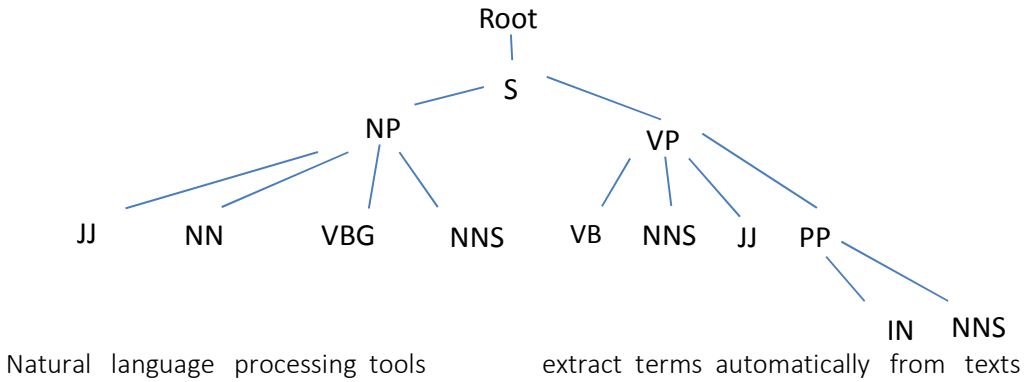

Fig. 1.9 - Arbre syntaxique de « natural langue processing tools extract terms automatically from texts »

1.3.2 Notre contribution : Architecture intégrante pour l'ingénierie des compétences

Plusieurs travaux ont appliqué certaines techniques de l'ingénierie des connaissances à l'ingénierie des compétences. Ils ont souvent utilisé un langage formel (logiques de description, théorie des ensembles, XML, *etc.)* pour définir une ontologie des compétences et effectuer du raisonnement sur ses composants. Toutefois, comme nous l'avons évoqué dans la section 1.2.4.1,

[6] NN : nom, NNS : nom au pluriel, VB : verbe au présent, VBG : gérondif, JJ : adjectif , IN : préposition, NP : groupe nominal, VP : groupe verbal, PP : groupe prépositionnel
[7] www.nltk.org/
[8] http://nlp.stanford.edu

dans ces travaux, les compétences ont été souvent confondues avec les concepts du domaine des compétences. Par conséquent, les compétences et ses c-ressources n'étaient pas distinguées. Ceci ne permet pas de faire du raisonnement sur les c-ressources des compétences. Par exemple, on ne peut pas inférer les c-ressources requises pour acquérir une compétence donnée, sachant qu'on possède une compétence proche de celle-là.

Pour identifier les compétences requises ou acquises (1er processus de l'ingénierie des compétences), les interviews sont souvent utilisées comme une méthode d'élicitation des connaissances.

Pour extraire des informations relatives aux compétences à partir des documents, plusieurs approches emploient des techniques de TAL, de fouille de textes ou d'apprentissage automatique.

Buitelaar&Eigner [2008] et Bordea *et al.* [2013] utilisent des patrons morphosyntaxiques pour identifier des compétences correspondant à des thèmes de recherche à partir des publications scientifiques.

D'autres auteurs utilisent la méthode LDA (Latent Dirichlet Allocation appartenant au groupe de méthodes factorielles) afin d'extraire conjointement des groupes homogènes d'auteurs et des thématiques à partir des collections d'articles [Mimno&McCallum, 2007], [Yuancheng *et al.* 2010]. Certains travaux utilisent l'annotation sémantique de documents à l'aide d'une ontologie du domaine et ensuite la recherche d'information pour identifier des compétences acquises [Sure *et al.* 2000], [Uren *et al.* 2006]. L'annotation sémantique est une technique qui associe à un objet (texte, image, *etc.*) ou à une partie de cet objet une description sémantique à l'aide d'une ontologie afin de bien le décrire [Prié&Garlatti, 2004].

D'autres construisent une ontologie du domaine à partir des documents du domaine considéré [Mika, 2005] ou ils enrichissent une ontologie existante en appliquant des techniques statistiques et de TAL aux documents de l'organisation en question ou de ses employés pour identifier des nouveaux concepts clés [Cameron *et al.* 2010] [Punnarut&Srihar, 2010].

Pour simplifier l'évaluation des compétences acquises (2ème processus d'ingénierie des compétences), certains travaux utilisent des méthodes d'extraction des connaissances à partir des documents associés aux individus, à leurs intérêts ou aux tâches réalisées. D'autres travaux utilisent la notion d'« intérêt » des employés à travers un système de recommandation [Lindgren *et al.* 2003]. Les « intérêts » sont considérés proches des compétences. L'auto-évaluation des compétences acquises guidée par une ontologie des compétences a été aussi proposée [Trichet *et al.* 2002].

Le recrutement et la formation sont les deux méthodes classiques reconnues pour l'acquisition des compétences (3ème processus d'ingénierie des connaissances). Les systèmes de recommandation peuvent aider à chercher des personnes adéquates pour une offre de recrutement à partir des bases de données ou des documents. Les systèmes de e-learning sont utilisés pour acquérir des nouvelles compétences, ils peuvent être aussi utilisés pour évaluer les compétences acquises, à l'aide des techniques de raisonnement [Baldoni, 2004], [Garro&Palopoli, 2003], [Colucci *et al.* 2005]. D'autres travaux utilisent des règles et/ou des mesures sémantique de similarité pour l'évaluation des compétences acquises [Blanchard *et al.* 2004] [Sure *et al.* 2000] [Laukkanen&Helin, 2005].

A partir des données du web, des travaux se sont intéressés à la recherche d'experts en tenant compte de l'hétérogénéité des documents et des sources [Aleman-Meza *et al.* 2007], [Stankovic et al. 2010], [Stankovic *et al.* 2011]. Ils ont montré notamment qu'il était possible d'agréger des informations hétérogènes à l'aide des technologies du web sémantique [Monaghan *et al.* 2010].

Enfin, pour établir une correspondance sémantique entre les profils requis et acquis (un des sous-processus principaux de l'exploitation des compétences) des travaux proposent des

algorithmes d'alignement (competence matching) basés sur des mesures sémantiques [Colucci *et al.* 2003b].

Toutefois, les différents travaux présentés ci-dessus se focalisent sur un seul type de processus parmi les quatre processus d'ingénierie des compétences.

Notre contribution

Nous avons proposé une architecture intégrante pour l'ingénierie des compétences (cf. Fig.1.10) [Berio&Harzallah, 2005a], [Berio&Harzallah, 2005b], [Berio&Harzallah, 2005c] [Berio&Harzallah, 2007]. Cette architecture définit les deux composants principaux d'un système d'ingénirie des compétences : une ontologie des compétences définie en se basant sur une ontologie de domaine, et une ontologie d'ingénierie des compétences (*i.e.* Competence mangement ontology). Elle identifie des relations entre ce système et d'autres éléments de son contexte : l'entreprise/organisation et les individus qu'on gère leurs compétences, les données à partir desquelles des compétences acquises ou requises peuvent être extraites, un système de gestion des formations pour acquérir des compétences, etc. Dans cette architecture, les compétences sont à représenter par une véritable ontologie à construire à partir de notre ontologie noyau des compétences (cf. section 1.2.4.2) offrant ainsi une modélisation unifiée des compétences pour les différents processus d'ingénierie des compétences (que ce soit pour les compétences acquises ou requises) et permettant d'y appliquer des techniques de raisonnement. L'ontologie d'ingénierie des compétences[9] permet de structurer et répertorier des méthodes et techniques d'ingénierie des connaissances selon les processus d'ingénierie des compétences et les ressources à partir desquelles les compétences acquises ou requises peuvent être identifiées. Les types de ressources suivants sont à intégrer dans cette ontologie : les historiques des formations reçues par des employés (et des candidats au recrutement) ; les données d'une organisation relatives à ses projets, ses activités ou ses tâches (se trouvant dans ses bases de données, ses documents ou les traces des activités réalisées) ; et les données des interviews. Cette ontologie permettra de structurer et répertorier également un ensemble de règles pour l'automatisation de l'identification et de l'évolution des compétences requises et acquises, en fonction du processus d'ingénierie des compétences et de son contexte. Nous pouvons citer des règles du processus d'entretien ; des « règles expertes » comme dans [Blanchard&Harzallah, 2004] et [Sure *et al.* 2000] ; des règles utilisant les historiques de formation des individus ; des règles utilisant les intérêts des individus [Lindgren *et al.* 2003] et des règles des méthodes d'évaluation comme dans un système de e-learning [Garro&Palopoli, 2003].

Nos modèles et notre architecture intégrante sont conformes aux conclusions citées dans [CedarCrestone, 2010] pour le développement des systèmes portant sur les ressources humaines. En effet, ils représentent un support pour le développement d'une façon intégrée et unifiée des fonctionnalités d'un système de gestion des ressources humaines et ils proposent un cadre complet pour son implémentation.

En 2005, j'ai participé au montage du STREP COMACO (Competence Management as a service for COllaborative work). COMACO est une proposition qui a porté sur le développement d'une plateforme pour l'ingénierie des compétences comme un service pour le travail collaboratif, basée sur notre architecture intégrante pour l'ingénierie des compétences. Dans cette proposition, nous avons prévu, entre autres, de construire l'ontologie d'ingénierie des compétences. Malheureusement cette proposition n'a pas été acceptée.

[9] Nous n'avons pas développé une ontologie d'ingénirie des compétences mais nous proposons l'utilisation de ce type d'ontologie dans un système d'ingénierie des compétences.

Nous avons suivi notre architecture intégrante pour : (1) l'enrichissement d'une ontologie des compétences (Master de V. Posea) [Posea, 2004], [Posea&Harzallah, 2004], (2) l'aide à l'évaluation des compétences acquises (Master de E. Blanchard), [Blanchard&Harzallah, 2004a], [Blanchard&Harzallah, 2004b], (3) la classification dynamique des compétences [Berio *et al.* 2007 ; 2008 ; 2011] et (4) l'extraction des compétences de CVs [Trichet *et al.* 2002]. Dans la suite, nous présentons le 2ème cas d'application de notre architecture.

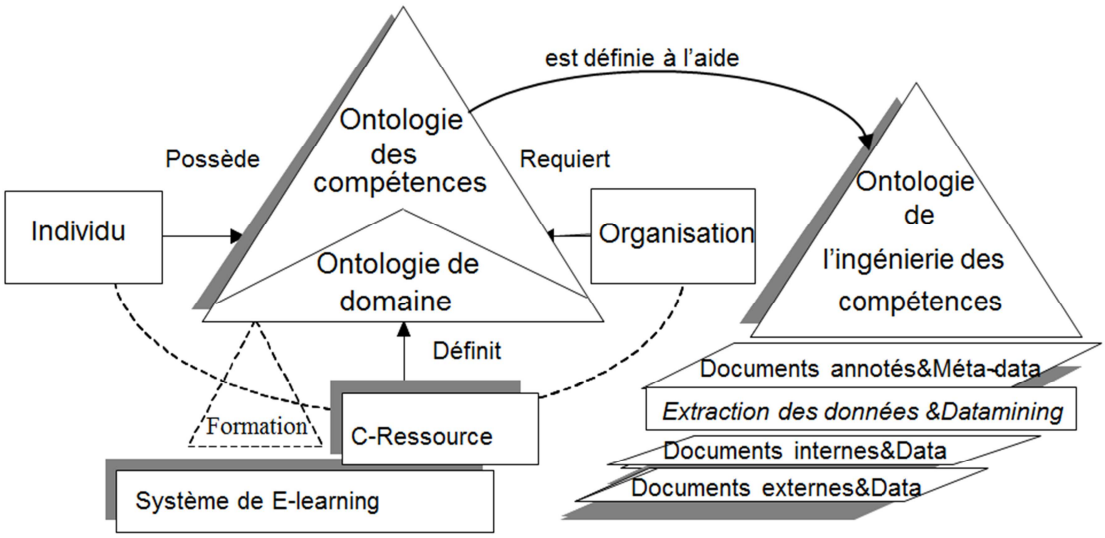

Fig. 1.10 – Architecture intégrante pour l'ingénierie des compétences ([Berio&Harzallah, 2007] avec modifications)

1.4 Conclusion

Dans ce chapitre, nous avons analysé les liens entre les notions de compétence, connaissance et donnée. Nous avons fait un bref état de l'art des modèles de représentation ainsi que des techniques pour leur ingénierie. Dans nos travaux, nous considérons que la modélisation et les techniques d'ingénierie sont deux manières pour définir, voire formaliser les liens entre ces trois concepts.

Nous avons présenté nos travaux de recherche sur l'articulation entre compétences et connaissances, en continuité avec ceux réalisés pendant notre thèse de doctorat. Plus précisément, nous avons présenté deux contributions majeures : (1) des modèles de connaissances (*i.e.* une ontologie noyau des compétences basée sur le modèle CRAI et le modèle CKIM pour une représentation intégrée des compétences et connaissances) et (2) une architecture intégrante pour l'ingénierie des compétences. Cette architecture se base sur une modélisation ontologique unifiée et fine des compétences permettant d'intégrer la réalisation des quatre processus principaux d'ingénierie des compétences. En plus, elle permet répertorier et structurer des techniques d'ingénierie des connaissances et les ressources associées pour l'extraction des compétences.

Nos travaux sur les compétences sont des travaux pionniers dans le domaine de l'ingénierie des compétences et ont été utilisés comme cadre pour de nouveaux travaux nationaux ou internationaux sur les compétences.

Comme perspective, cette architecture pourrait s'étendre à une architecture intégrante pour l'ingénierie des compétences, des connaissances et des données (Fig. 1.11). En effet, en plus des quatre processus d'ingénierie des compétences, d'autres processus d'une organisation qui requièrent des connaissances pourraient être rajoutés à cette architecture . Par exemple, on peut s'intéresser au processus de marketing d'une organisation et déterminer les besoins potentiels des clients de cette organisation ou leur degré de satisfaction en analysant leurs avis sur ses produits. On peut s'intéresser aussi aux processus de maintenance préventive et analyser les données relatives aux comportements des machines et prédire leurs pannes. D'autres types de données (*e.g.* données structurées, traces) et d'autres techniques pour extraire les connaissances nécessaires pour un type de processus sont à rajouter dans notre architecture pour couvrir l'ingénierie des connaissances pour les différents processus d'une organisation [Abiteboul *et al.* 2017], [André *et al.* 2017]. Enfin, dans cette architecture, nous mettrons l'accent sur les liens bidirectionnels entre les connaissances et données, via son ontologie d'ingénierie des connaissances. En effet, les techniques qui requièrent certains types de données pour extraire des connaissances seront identifiées, mais aussi celles qui requièrent des connaissances pour pouvoir traiter des données.

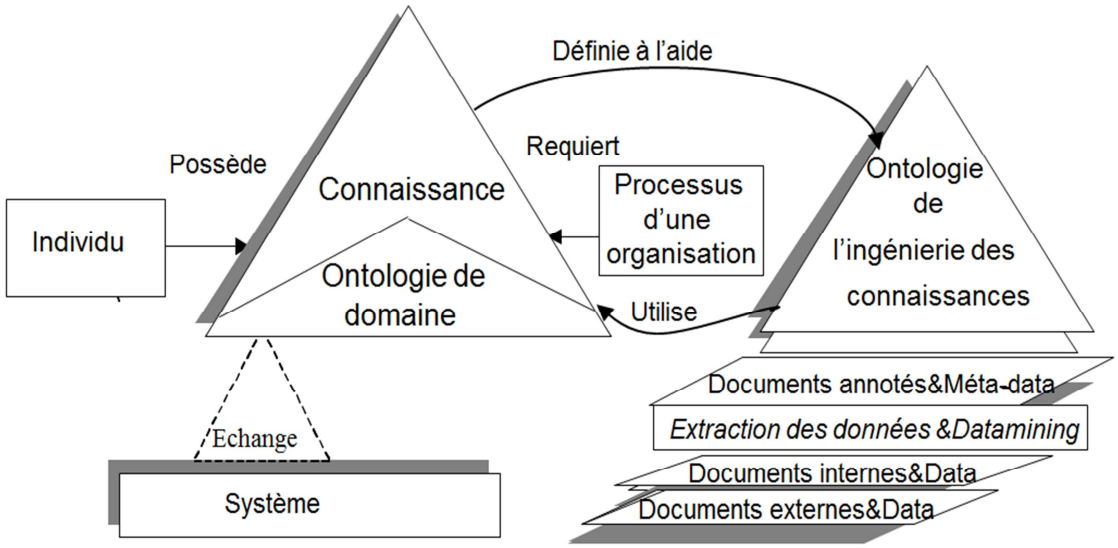

Fig. 1.11 - Vers une architecture intégrante pour l'ingénierie des compétences, connaissances et données

Par ailleurs, notre architecture intégrée relie bien nos différents travaux de recherche (cf. Fig. 1 du chapitre Introduction). Dans les chapitres suivants, nous reconsidérons certaines méthodes d'ingénierie des connaissances, présentées dans ce chapitre. Dans le chapitre 2 et le chapitre 3, nous nous intéressons aux techniques d'ingénierie des connaissances pour la construction et la validation semi-automatiques d'ontologie à partir de textes. Le chapitre 4 porte sur les mesures sémantiques appliquées à une ontologie pour définir la sémantique des données (*e.g.* objet, terme, gène) et générer des nouvelles connaissances pour la classification, l'annotation, la recherche d'information...

Chapitre 2 : De la construction manuelle d'ontologie vers la construction semi-automatique

Sommaire

2.1 Introduction

Etroitement associées à l'essor des technologies du web sémantique, les ontologies sont devenues un des modèles majeurs de représentation des connaissances dans des domaines très variés. Un des challenges du web sémantique est de les utiliser pour relier entre elles des masses de données hétérogènes [Gruninger&Obrst, 2014].

Selon les besoins et les domaines d'application, on peut utiliser une ou plusieurs ontologies existantes, en totalité ou en partie.. On peut aussi avoir besoin d'en construire une nouvelle qui s'adapte aux besoins d'une application et de son domaine. Les ressources à partir desquelles une ontologie est construite peuvent être diverses et variées. Plusieurs méthodes de construction sont développées pour s'adapter à ces différentes ressources. De plus, une ontologie peut être construite manuellement à partir des connaissances des experts ou de ressources documentaires. Les ressources documentaires utilisées dans ce cas doivent être de nombre raisonnable et provenir de sources fiables. L'étape de conceptualisation d'ontologie est une des étapes du processus de construction d'ontologie les plus lourdes et longues à réaliser. Des méthodes ont été proposées pour sa réalisation, définissant des principes et règles de conceptualisation. Cependant, la réalisation de cette étape, reste complexe, notamment quand elle concerne un projet de grande envergure. Dans le cas de l'utilisation d'une masse importante de documents, le recours à l'automatisation (ou la semi-automatisation) de la conceptualisation d'ontologie devient indispensable. Pour ce faire, des techniques de disciplines variées ont été proposées : techniques de TAL, de fouille de textes, d'apprentissage automatique ou de raisonnement logique [Buitelaar *et al.* 2005]. Leurs résultats sont prometteurs mais ne sont pas de qualité suffisante pour être exploitable. Des travaux de recherche sont toujours en cours pour améliorer ces techniques ou en développer des nouvelles, afin d'obtenir des ontologies de bonne qualité.

Dans ce chapitre, nous présentons tout d'abord des approches existantes pour la construction d'ontologie et leur évolution par rapport à la diversité et à l'évolution des ressources à partir desquelles ces ontologies pourraient être construites (section 2.2). Nous nous focalisons ensuite sur l'étape de conceptualisation d'ontologie et nous montrons en particulier que les approches existantes pour sa réalisation restent générales et insuffisantes pour obtenir un résultat intéressant, surtout quand l'ontologie à construire porte sur un domaine vaste et complexe (section 2.3).

Dans la section 2.4, nous nous focalisons sur les approches semi-automatiques de conceptualisation d'ontologie à partir de textes. Tout d'abord, nous présentons un état de l'art des techniques existantes pour l'automatisation des différentes tâches de conceptualisation d'ontologie (section 2.4.1). Nous comparons ensuite des approches et outils de la littérature pour la conceptualisation semi-automatique d'ontologie à partir de textes, en utilisant un cadre structurant que nous avons développé (section 2.4.2). Nous présentons brièvement dans la section 2.4.3 les améliorations apportées à l'outil Text2Onto pour la construction automatique d'ontologie à partir de textes en anglais et son adaptation à la langue française.

Dans la dernière partie, nous rapportons notre expérience en construction d'ontologie dans le cadre des projets UEML, ISTA3 et KIFANLO auxquels nous avons participés, les difficultés que nous avons rencontrées et le rôle important qu'une ontologie noyau a joué pour construire une ontologie (section 2.5). Cette expérience nous a orienté vers le développement d'une approche semi-automatique de construction et validation intégrées d'ontologie basée sur une ontologie noyau formelle. Nous présentons dans le chapitre suivant les grandes lignes de cette approche qui représente une des perspectives de notre projet de recherche.

2.2 Evolution des méthodes de construction d'ontologie

Depuis plus de 25 ans, l'ingénierie des connaissances s'intéresse aux ontologies et à leur construction. Des approches générales ascendantes, descendantes ou mixtes ont été proposées. Elles sont souvent similaires et proches à celles de construction d'une application informatique ou d'un modèle. Elles comprennent souvent les grandes étapes suivantes : (1) la spécification de l'ontologie à construire, cette étape concerne souvent la définition du domaine de l'ontologie, ses compétences et ses scénarii d'usage ; (2) sa conceptualisation qui comprend l'acquisition des connaissances et la spécification informelle de ses composants ; (3) sa formalisation avec un langage formel, souvent basée sur une logique de description et permettant de faire des raisonnement ; (4) son implémentation en choisissant un langage (*e.g.* OWL) et un éditeur d'ontologie (*e.g.* Protégé) ; (5) sa validation ; et (6) son évolution.

Certaines approches s'intéressent à la construction d'ontologie en général et pourraient s'adapter à une construction manuelle ou automatique, *e.g.* MethOntology [Gomez-Perez *et al.* 2001]. D'autres approches sont plutôt adaptées au développement d'une ontologie pour une application spécifique. Elles mettent l'accent sur la première étape du processus de construction pour s'assurer de la faisabilité de l'ontologie à développer, *e.g*. On-To-Knowledge [Staab *et al.* 2001].

Certaines étapes de ce processus, complexes et à effectuer tout au long de la construction, sont devenues des processus d'aide, réalisés en parallèle à la construction proprement dite, *e.g.* l'étape d'acquisition des connaissances ou l'étape de validation. Enfin, des processus de gestion permettant de bien planifier la construction ou la documentation.

Le développement d'ontologies pour des applications ou projets réels a impliqué de considérer des ontologies plus grandes dont la construction nécessite l'intervention de plusieurs partenaires. Ceci a encouragé la proposition de méthodologies collaboratives pour l'aide à la construction, la validation ou l'enrichissement d'ontologie, *e.g.* DILIGENT [Pinto *et al.* 2004], Web-Protege [Tudorache *et al.* 2008], GOSPL [Debruyne *et al.* 2013], UPON Lite [De Nicola&Missikoff, 2016]. Les deux dernières méthodologies mettent l'accent sur le rôle primordial de l'expert du domaine de l'ontologie. En outre, afin de réduire la durée de ce processus, De Nicola&Missikoff [2016] recommandent de faire construire une ontologie par les experts du domaine et de ne faire intervenir un expert des ontologies que dans la phase de sa formalisation.

Avec l'accessibilité croissante des données et des textes, l'intérêt de la construction semi-automatique d'ontologie à partir de textes est devenu incontestable. Des méthodologies et des plateformes ont été proposées pour soutenir cette construction.

Les méthodologies Terminae [Aussenac-Gilles *et al.* 2008], [Szulman, 2013] ou Dafoe [Charlet *et al.* 2010] font coopérer un ensemble d'outils et s'appuient sur une plateforme technique pour concevoir une ontologie. Leur cadre méthodologique est constitué de quatre étapes : (1) la constitution d'un corpus de textes, (2) l'analyse linguistique de ce corpus, (3) la conceptualisation, et (4) la formalisation de l'ontologie obtenue.

La méthodologie NeOn [Suarez-Figueroa *et al.* 2008], [Suarez-Figueroa, 2010] offre une plateforme d'outils d'extraction automatique de connaissances, de construction et d'évaluation d'ontologie, de mapping d'ontologies, *etc*. Elle a été proposée dans un cadre plus large répondant au paradigme des données liées, des ontologies pour le web sémantique, et de la multitude d'ontologies pouvant exister dans le web [Suárez-Figueroa *et al.* 2012]. Elle recommande pour cela :

– la construction d'ontologie basée sur la réutilisation, la restructuration, la modification et l'adaptation des différentes sources de connaissances déjà existantes ;

– la réutilisation dynamique dans une ontologie des concepts définis dans d'autres ontologies disponibles en ligne ;
– le rôle de plus en plus important de la collaboration et de l'argumentation durant la construction d'ontologie.

La prolifération d'ontologies sur le web a poussé vers la construction de réseaux d'ontologies : des ontologies interconnectées et interdépendantes sur le web. La méthodologie NeOn a été étendue pour la construction de ce type de réseau [Suárez-Figueroa, 2010]. Certains travaux ont cherché à formaliser les liens entre les ontologies qui composent ces réseaux [Rohrer, 2012].

D'autres travaux ont préconisé le développement et l'utilisation d'ontologies « lights » pour annoter les métadonnées des documents du web afin de les relier. L'ontologie BFO (Basic Formal ontology) est une des ontologies « lights » du web des données [Bittner&Smith, 2004], [Grenon *et al.* 2004]. Elle est composée de concepts généraux communs à plusieurs domaines. Elle a été utilisée aussi pour décrire des concepts des réseaux d'ontologies. D'autres ontologies « lights » mais plus spécifiques à certains domaines ont été développées, par exemple pour décrire le domaine des réseaux sociaux (*e.g.* l'ontologie FOAF[10] ou celui de l'internet des objets e.g. SAREF[11]). DBpedia[12] a été construite afin de relier les données de Wikipédia à d'autres sources du web.

Dans nos travaux de recherche, nous nous sommes intéressés plus particulièrement à deux étapes du processus de construction d'ontologie : l'étape de conceptualisation et l'étape de validation. Ce sont les deux étapes du processus de construction les plus longues et complexes à réaliser. Nous avons cherché à améliorer leur réalisation manuelle ou semi-automatique. Enfin, une de nos perspectives de recherche est de proposer une méthodologie semi-automatique pour les réaliser à partir d'une masse importante de données. Dans ce chapitre, nous nous focalisons sur l'étape de conceptualisation. L'étape de validation est l'objet du chapitre suivant.

2.3 Conceptualisation d'ontologie

Plusieurs travaux se sont intéressés à l'étape de conceptualisation d'ontologie et ont défini ses sous-étapes et des méthodes, des directives ou des principes pour sa réalisation. MethOntology décompose cette étape en sept tâches [Fernandez *et al.* 1997], [Corcho *et al.* 2005], [Gomez-Perez *et al.* 2001]. La première tâche est la construction d'un glossaire de termes faisant référence à des connaissances pouvant faire partie de l'ontologie à définir. La deuxième tâche est la construction de taxonomies de concepts (les termes synonymes du glossaire sont associés à un même concept et les concepts partageant les mêmes propriétés sont regroupés par catégorie afin de construire une taxonomie). La troisième tâche est la construction des relations binaires ad-hoc. Une relation relie deux concepts identifiés dans la tâche précédente. La quatrième tâche est la définition d'un dictionnaire de concepts. Toutes les informations relatives aux concepts (leur sémantique, leurs attributs, leurs instances, *etc*.) sont regroupées dans le dictionnaire de concepts. La cinquième tâche est la description des relations ad-hoc, des attributs et des constantes. Pour chaque relation identifiée dans la tâche 3, son étiquette, les concepts qu'elle relie, sa cardinalité, ses relations inverses, *etc.* sont définis. Dans, la sixième tâche, les axiomes de l'ontologie à définir sont identifiés et décrits formellement. La septième tâche est la description des instances.

[10] http://www.foaf-project.org/
[11] https://sites.google.com/site/smartappliancesproject/ontologies
[12] http://wiki.dbpedia.org/

Pour chaque instance mentionnée dans le glossaire de termes, son nom, les concepts qui lui sont associés et les valeurs de ses attributs sont précisés.

Plusieurs ontologies ont été construites en suivant cette méthodologie [Gomez-Perez *et al.* 2001], [Corcho *et al.* 2005]. La plupart des auteurs considèrent que son mérite principal est d'avoir bien décomposé et explicité chacune des tâches de la conceptualisation de façon qu'elle puisse s'adapter à une construction manuelle ou semi-automatique et à partir ou non de textes.

D'autres approches ont aussi détaillé et formalisé la conceptualisation ou ont amélioré la réalisation d'une ou de plusieurs de ces tâches. Par exemple, l'approche Terminae détaille le passage de termes à concepts et celle de la formalisation des concepts [Aussenac *et al.* 2008].

Des principes généraux ont insisté de définir clairement les composants de l'ontologie [Gruber, 1993], [Guarino, 1998]. Bachimont *et al.* [2002] ont proposé dans la méthode ARCHONTE, quatre principes pour bien oragniser des concepts en une taxonomie. argumenter le rajout d'une relation de subsumption. Guarino&Welty [2000] ont défini dans la méthodologie OntoClean, utilisant des règles pour aider à vérifier ce type de relation, mais aussi pour décider si une donnée est un concept, un attribut ou une relation. Cette méthodologie se base sur l'affectation de quatre méta-propriétés (*i.e.* identité, rigidité, unité et dépendance) aux concepts d'une ontologie. La cohérence logique de cette ontologie est vérifiée ensuite par rapport aux règles liées à ces méta-propriétés [Guarinon&Welty, 2002]. Cependant, l'affectation de ces méta-propriétés devient rapidement une tâche fastidieuse, voire complexe [Hernandez, 2006], [Pomohaci, 2006].

Il existe aussi des patrons pour la conception d'ontologie (ODP[13] : Ontology Design Pattern) proposés dans NeOn pour bien modéliser certains cas de conception bien précis et récurrents. De nouveaux patrons sont récemment proposés et classés en six catégories [Hitzler *et al.* 2016] : (1) patrons de structures (structural ODP) pour bien formaliser, généralement en OWL, certaines situations par exemple la notion de partition ; (2) patrons de correspondances (correspondance ODP) pour aider à transformer un modèle qui n'est pas une ontologie en une ontologie ; (3) patrons portant sur le contenu (content ODP) qui proposent une modélisation de certains concepts appartenant à un domaine générique (upper domain) ou à un domaine spécifique ; (4) patrons lexico-syntaxiques pour identifier des relations terminologiques entre termes à partir de textes, ces patrons sont évidemment similaires à ceux présentés dans la section 2.4.1 ; (5) patrons de description (Description ODP) pour définir des meilleures pratiques pour nommer les artefacts d'une ontologie et les annoter ; et (6) patrons de raisonnement afin d'obtenir certains résultats de raisonnement en fonction du raisonneur utilisé. D'autres patrons ont été utilisés dans le framework OntoCase pour l'enrichissement d'ontologie [Blomqvist, 2009]. L'exploitation des patrons ODP semble très intéressante, notamment quand on connait le patron adéquat et le moment de son utilisation dans un processus de conceptualisation d'ontologie. Toutefois, les patrons portant sur le contenu, pourraient ne pas correspondre nécessairement au domaine d'une ontologie à développer. Enfin, les patrons ODP portent en grande partie sur la formalisation d'une ontologie plus que sur sa conceptualisation.

Par ailleurs, dans un contexte où les ontologies portent sur des domaines vastes et multi-disciplinaires, et afin d'aider à les construire, les enrichir, les utiliser ou les maintenir, leur modularisation est recommandée [Gangemi *et al.* 2006]. Elle consiste à (ré)organiser l'ontologie en plusieurs fragments (*i.e.* modules) autonomes, indépendants et réalisables [d'Aquin, 2012], [Ben Mustapha *et al.* 2013]. Plusieurs travaux se sont intéressés à la notion de module dans une ontologie (notion reconnue depuis longtemps comme un élément structurant et incontournable

[13] http://ontologydesignpatterns.org

des artefacts logiciels). Certains d'entre eux ont proposé l'utilisation d'une ontologie noyau pour construire une ontologie modulaire [Gangemi&Borgo, 2004], [Kutz&Hois, 2012], [Gaoussou *et al.* 2014]. Par exemple, Gaoussou *et al.* [2014] ont utilisé deux ontologies noyaux (i.e. une ontologie noyau des maladies infectieuses et une ontologie noyau de la propagation des maladies infectieuses) pour construire une ontologie de la schistosomiase (IDOSCHISTO). D'autres travaux alignent une ontologie noyau ou une ontologie de haut niveau à une ontologie du domaine pour mieux définir les concepts de cette dernière et lui imposer une structure prédéfinie [Gangemi *et al.* 2002]. Par exemple, Després et Szulman [2007] ont aligné des micro-ontologies du droit à une ontologie noyau du droit ; Burita *et al.* [2012] ont mappé l'ontologie noyau NEC (Network Enabled Capabilities) à l'ontologie du domaine de NEC. Dans nos travaux, nous considérons qu'une ontologie noyau d'un domaine est composée du minimum de concepts et de relations (nommés concepts et relations noyaux) nécessaires pour décrire ce domaine [Guarino *et al.* 2009], [Burita *et al.* 2012]. Nous supposons que tout concept d'une ontologie de ce domaine construite à partir de cette ontologie noyau est une spécialisation d'un de ces concepts noyaux. Nous supposons aussi, que chaque relation non taxonomique entre deux concepts est une spécialisation d'une relation noyau. D'autres relations entre les sous-concepts d'un même concept noyau peuvent être rajoutées.

En conclusion, plusieurs travaux ont traité l'étape de conceptualisation d'ontologie afin de la structurer et améliorer sa réalisation. Il nous semble quand-même que certaines questions liées à la conceptualisation proprement dite et non à la formalisation restent ouvertes et il n'est pas facile d'y répondre, surtout quand on aborde la conceptualisation d'une ontologie de grande taille. Nous pouvons en citer quelques-unes, que nous avons posées au sein du groupe de conceptualisation de l'ontologie KIFANLO ou de l'ontologie ISTA3 : (1) Selon quels concepts faut-il structurer une ontologie, si plusieurs structurations sont possibles ? (2) Comment peut-on décider rapidement si un terme est pertinent pour une ontologie et choisir sa modélisation comme un concept, une instance, une propriété d'un concept ou une relation entre deux concepts ? (3) Comment peut-on décider rapidement qu'un concept est une spécialisation d'un autre concept ?

Dans les projets KIFANLO et ISTA3, pour nous aider à répondre à ces questions lors de l'étape de conceptualisation, nous nous sommes focalisés sur l'objectif de l'ontologie à développer, d'une part, et d'autre part nous avons utilisé une ontologie noyau. Nous sommes convaincus qu'une ontologie noyau pourrait bien guider la réalisation de cette étape, notamment si l'ontologie noyau utilisée comprend un nombre suffisant d'axiomes qui contraignent la définition de ses artefacts. En dernière section de ce chapitre, nous discutons de nos expériences en construction d'ontologie avec une ontologie noyau et nous présentons dans le chapitre suivant notre contribution pour la construction et la validation d'ontologie basée sur une ontologie noyau.

Dans la suite, nous nous intéressons plus particulièrement à la conceptualisation semi-automatique d'ontologie et nous discutons des techniques, méthodes et outils pour la réaliser. Nous considérons l'acquisition des connaissances comme une partie intégrante de la conceptualisation. Nous ne traitons pas les activités de choix et d'analyse de corpus qui précèdent l'acquisition des connaissances. Toutefois, ce sont deux activités clés dans la construction d'ontologie [Lame, 2002], [Bourigault *et al.* 2004], [Condamines, 2005].

2.4 Conceptualisation semi-automatique d'ontologie

Les ontologies étaient construites initialement à la main en se basant bien souvent sur les savoir-faire des experts du domaine. Leur popularisation et l'accès facile aux ressources textuelles ont incité au passage à une construction semi-automatique. La construction (et plus

particulièrement la conceptualisation) semi-automatique est un processus comprenant plusieurs tâches complexes en elles-mêmes (*e.g.* l'extraction des termes et relations terminologiques, la définition des concepts). Malgré la difficulté de ce challenge, la nécessité du recours à la construction semi-automatique a stimulé fortement le développement de méthodes et outils variés.

Les outils logiciels les plus utilisés sont certainement les éditeurs d'ontologie, qui offrent une palette d'outils facilitant la conceptualisation. Parmi les éditeurs les plus performants on peut citer Protégé[14], NeOn (neon-toolkit.org) et TopBraid Composer (www.topquadrant.com).

Au-delà des éditeurs, d'autres recherches se sont orientées depuis plus que 15 ans vers la conceptualisation automatisée d'ontologie à partir d'un corpus textuel [Faure &Nédellec, 1998], [Nazarenko&Hamon, 2002], [Velardi *et al.* 2005], [Buitelar *et al.* 2005], [Jinsoo *et al.* 2011], [Velardi *et al.* 2013], [Haidar-Ahmad *et al.* 2016]. Dans les approches développées, souvent, une ou plusieurs tâches du processus de conceptualisation sont réalisées à l'aide d'une ou plusieurs techniques de TAL, de fouille de textes ou d'apprentissage automatique.

Nous présentons dans la suite les techniques existantes pour réaliser une ou plusieurs tâches de l'étape de conceptualisation d'ontologie et les méthodes et outils proposés pour les accomplir.

2.4.1 Techniques de conceptualisation semi-automatique d'ontologie

Plusieurs techniques ont été adaptées et d'autres ont été développées pour automatiser les différentes tâches de conceptualisation d'ontologie. Nous avons choisi de considérer les tâches définies dans le processus de conceptualisation dans MethOntology comme une référence pour classifier et comparer ces techniques [Gherasim *et al.* 2013]. Nous considérons particulièrement les trois premières tâches : la tâche de construction d'un glossaire de termes, la tâche de construction de taxonomies de concepts et la tâche de construction des relations ad-hoc.

2.4.1.1 Techniques pour la construction d'un glossaire de termes

Deux types de techniques existent pour l'extraction de termes à partir de textes : les techniques statistiques et les techniques linguistiques [Buitelar *et al.* 2005], [Pazienza *et al.* 2005]. Les techniques statistiques utilisent souvent la fréquence d'un terme (tf), sa fréquence relative (tf-idf) dans un document d'un domaine ou les fréquences de ses cooccurrences avec d'autres termes clés. Elles visent à déterminer les termes pertinents pour un domaine donné. Les techniques linguistiques se basent sur une analyse morphosyntaxique d'un texte et l'utilisation de patrons pour extraire des termes. Par exemple, on extrait de la phrase « natural language processing tools extract terms automatically from texts »[15] les termes « natural language processing tool», « natural language », « language », « tool », « term », et « text » en utilisant respectivement les patrons morphosyntaxiques « JJ/NN/VBG/NN[16] », « JJ/NN » et « NN ». Souvent les techniques linguistiques et statistiques sont utilisées conjointement, formant ce qu'on appelle des techniques mixtes ou hybrides.

Pour l'attribution d'une définition à un terme, Navigli&Velardi [2004 ; 2005] ont proposé des stratégies pour extraire cette définition d'une façon semi-automatique à partir du web.

2.4.1.2 Techniques pour la construction de taxonomies de concepts

Le passage de termes à concepts consiste à identifier des relations de synonymie entre termes et de choisir un label pour chaque classe sémantique (i.e. un ensemble de termes synonymes)

[14] protege.standford.edu
[15] Cf. section 1.3.1.3
[16] JJ : adjectif, NN : nom, VBG : gérondif

ayant un sens unique et représentant un concept. Les techniques qui se base sur l'analyse distributionnelle des termes (cf. section 1.3.1.2) peuvent aider à la détermination de termes synonymes. Par exemple, dans l'article de Navigli&Velardi [2004] « tool » et « system » ont été identifiés plusieurs fois comme le sujet des verbes « extract from », « extract » et « use », ce qui pourrait suggérer de les regrouper en une classe de termes synonymes. Ces techniques ne sont pas vraiment précises pour l'identification des termes synonymes. En effet, elles déterminent des termes proches sémantiquement mais pas nécessairement synonymes.

Une ressource externe peut être utilisée pour déterminer des termes synonymes, « synsets » de WordNet par exemple. Cependant, on se confronte souvent au problème de polysémie des termes.

Pour l'extraction de relations taxonomiques lexicales (*i.e.* relations *est-un* entre termes) cinq techniques sont présentées dans la littérature : les techniques du type « document subsumption », les techniques structurelles, les techniques contextuelles, l'utilisation d'une ressource externe, et les techniques hybrides qui sont des combinaisons des techniques précédentes [Gherasim *et al.* 2013b]. Les techniques de type « document subsumption » sont basées sur l'idée qu'un terme est une spécialisation d'un autre terme si le deuxième apparait dans tous les documents où apparait le premier [Sanderson&Croft, 1999]. Les techniques structurelles utilisent la structure interne d'un terme (*e.g.* le terme « ontologie noyau » *est-un* « ontologie » ou « domain ontology » *est-un* « ontology »). La structure d'un texte et notamment ses ponctuations peuvent aussi être utilisées pour identifier des relations *est-un* [Kamel&Rothenburger, 2011]. Par exemple, en considérant « : » dans la première phrase du paragraphe suivant, on peut extraire que « une technique basée sur une analyse distributionnelle » *est-un* « une technique contextuelle » et que « une technique basées sur les patrons » *est-un* « une technique contextuelle ».

Concernant les techniques contextuelles, nous pouvons en distinguer deux familles : (1) les techniques basées sur l'analyse des distributions des termes et (2) les techniques basées sur les patrons. Les techniques de la première famille se basent sur la fréquence de co-locations de deux termes (i.e. termes qui apparaissent dans le même contexte) pour déterminer une proximité sémantique entre eux. Un contexte d'un terme peut être défini par un document, un paragraphe, une phrase, *etc*. où ce terme apparaît, ou par un attribut, une relation ou un contexte syntaxique qu'il possède. Ensuite, trois techniques peuvent être appliquées sur les distributions des termes pour déterminer des classes de termes[17] et des relations *est-un* entre elles : (1) une mesure de similarité ou d'inclusion pour comparer deux termes en fonction des contextes qu'ils partagent (*e.g.* la mesure de Dice ou celle de Jaccard appliquées aux ensembles de contextes de chaque terme) et une méthode de classification non-supervisées (e.g. k-means) ou supervisées (*e.g.* k plus proches voisins) [Lenci&Benotto, 2012], [Roller *et al.* 2014], (2) des techniques de classification basées sur les treillis de Galois (l'analyse formelle des concepts (FCA) et l'analyse relationnelle de concepts (RCA)) [Buitelaar *et al.* 2005], [Toussaint, 2011], [Mondary, 2011] et (3) des techniques de recherche de motifs fréquents (séquentiels ou non) et d'extraction des règles d'association [Maedche&Staab, 2000]. Nous présentons ce dernier type de techniques dans la section 2.4.1.3 parce qu'il s'applique souvent à l'extraction des relations ad-hoc.

Dans la deuxième famille de techniques contextuelles, l'utilisation de patrons pour l'identification des relations *est-un* implique le recours à des patrons existants ou à l'apprentissage de nouveaux patrons. Plusieurs patrons ont été proposés dans la littérature, les plus connus sont les patrons de Hearst ou les patrons proposés dans [Ogata&Collier, 2004]. Ces patrons sont connus par une précision acceptable et un faible rappel [Panchenko *et al.* 2016]. La

[17] Il ne s'agit pas nécessairement de classes sémantiques

qualité de leurs résultats dépend aussi de la qualité de la phase d'étiquetage morphosyntaxique et du type du corpus utilisé (*e.g.* glossaire ou littéraire) [Aussenac-Gilles, 2005]. Pour améliorer la qualité des patrons de Hearst, d'autres travaux en ont proposé plusieurs variétés prenant en compte la diversité de formulation de la relation d'hyperonymie [Seitner *et al.* 2016]. Nous approfondissons cette technique dans la section 2.4.1.3. Les approches basées sur l'apprentissage de patrons ont un rappel plus élevé que celui des précédentes [Pantel *et al.* 2004], [Sheena *et al.* 2016]. Cependant, elles souffrent du fléau de la dimension [Shwartz *et al.* 2016].

D'autres approchent prédisent des relations d'hyperonymie entre termes en se basant sur l'analyse distributionnelle de leurs contextes. Elles souffrent généralement du fléau de la dimension. Récemment, la technique de « *word embeddings*» *est utilisée pour diminuer* le nombre de dimensions pour représenter des termes et améliorer ainsi la performance des approches basées sur l'analyse distributionnelle [Mikolov *et al.* 2013], [Shwartz *et al.* 2016].

Enfin, des approches mixtes ont été proposées combinant l'utilisation de patrons et les approches basées sur l'analyse distributionnelle [Schropp *et al.* 2013].

2.4.1.3 Techniques pour l'extraction de relations ad-hoc

L'extraction de relations lexicales ad-hoc est le plus souvent basée sur l'analyse linguistique de texte et l'utilisation de patrons morphosyntaxiques. Parfois, des sources de connaissances externes ou des techniques basées sur l'analyse distributionnelle sont aussi utilisées. Ces dernières se basent sur la fréquence de cooccurrences de deux termes pour déterminer une connectivité sémantique entre eux. Cependant, il est difficile d'interpréter cette connectivité sémantique et de lui associer une relation. Par exemple, dans l'article [Navigli&Velardi, 2004], les termes « tool » et « text » coexistent dans plusieurs phrases, mais on ne peut pas automatiquement identifier la sémantique de la relation entre eux.

Les approches probabilistes comme les techniques de motifs fréquents (séquentiels ou non), les règles d'association ou les réseaux bayésiens ont montré leur intérêt pour traiter des distributions et extraire des relations avec une certaine incertitude. Les techniques de recherche des motifs fréquents et l'extraction des règles d'association sont souvent utilisées conjointement pour déterminer des relations entre termes, préalablement identifiés [Cimiano *et al.* 2005]. En effet, un motif est souvent un tuple de termes [Di Jorio *et al.* 2007] ou d'attributs de termes [Bendaoud *et al.* 2005]. Entre les termes (ou les attributs de termes) d'un motif fréquent, on établit des règles d'association avec une certaine probabilité, qui pourraient aider à déterminer des relations étiquetées ou non entre eux. Par exemple, un motif séquentiel pourrait être un triplé de type (NP1, Verbe, NP2) et il est fréquent s'ils apparait plusieurs fois dans un corpus. La règle d'association déterminée à partir d'un motif fréquent de ce type est que le couple de termes (NP1, NP2) est lié par une relation ayant le label du verbe de ce motif.

Les réseaux bayésiens sont souvent appliqués à des individus ou à des instances de leur attributs afin d'extraire des relations causales entre ces individus ou entre leurs attributs. Ces relations pourraient enrichir une ontologie et étendre sa modélisation vers une modélisation probabiliste [Ding & Peng, 2004], [BenIshak *et al.* 2011]. Par exemple, si on considère les trois concepts « ontology learning tool », « technique » et « ontology artefact », et les deux relations Extract(ontology learning tool, ontology artefact) et Use(ontology learning tool, technique). Si on détermine qu'il y a une probabilité forte qu'une instance de « ontology learning tool » liée à une instance de « ontology artefact » soit aussi liée à une instance de « technique », on pourrait suggérer de rajouter une nouvelle relation entre « technique » et « ontology artefact ». Toutefois, les approches probabilistes requièrent généralement un nombre très important de données pour les appliquer.

Les patrons morphosyntaxiques sont utiles pour l'indentification des relations communes à plusieurs domaines, comme par exemple la relation de composition ou la relation de causalité. Pour des domaines spécifiques, il est nécessaire de définir des patrons adéquats qui s'adaptent au domaine et au type de corpus de textes considérés. Par exemple, on peut utiliser le patron « **/NN[18]/**/extract(VB)/**/«from»/**/NN »[19] pour identifier la relation ExtractFrom(tool, text) à partir de la phrase « natural language processing tools extract terms automatically from texts » (cf. section 1.3.1.3). Cependant, vue la difficulté de connaitre préalablement toutes les relations spécifiques d'un domaine, des patrons constitués seulement d'étiquetages morphosyntaxiques ont été proposés. Par exemple, le patron « **/NN/**/VB/**/NN » appliqué à la phrase précédente permet d'identifier quatre relations : Extract(language, term), Extract(tool, term), Extract(language, text) et Extract(tool, text) et le patron « **/NN/**/VB)/**/«from»/**/NN » identifie deux relations : Extractfrom(language, text) et Extractfrom(tool, text). Cette technique peut donner un nombre important de relations avec un rappel et une précision faibles. En général, l'utilisation de patrons de surface (*i.e.* un patron linguistique qui ne prend pas en compte des dépendances syntaxiques) a ses limites. L'absence d'informations linguistiques profondes dans les patrons de surface est une source d'erreurs, *e.g.* Extractfrom(language, term). Ceci est à cause d'une part du fait que les phrases d'un texte ont parfois une structure linguistique non valide ou complexe ; et d'autre part du fait qu'elles peuvent comprendre des informations inutiles sémantiquement.

Des techniques d'apprentissage de patrons utilisant des phrases d'entrainement qui expriment la relation à identifier, ont été utilisées. Dans ces techniques, la structure de ces phrases est analysée pour identifier des patrons pour la relation recherchée [Faure&Nedellec, 1998a]. Certaines approches d'apprentissage de patrons sont basées sur une analyse linguistique d'un corpus d'apprentissage et sur la fouille de motifs séquentiels [Szathmary, 2006]. De plus, elles utilisent une liste de contraintes pour réduire le nombre important de motifs séquentiels extraits [Cellier *et al.* 2010], [Béchet *et al.* 2012]. Ces approches se réduisent souvent à l'identification des relations connues entre des entités nommées identifiées préalablement. Elles ont une précision intéressante et un rappel moyen parce qu'elles ne prennent pas en compte les relations syntaxiques entre les items d'un motif. Pour pallier aux limites des patrons de surface, certaines approches d'apprentissage représentent chaque phrase d'entrainement par un arbre de dépendances afin de déterminer des patrons utilisant des informations syntaxiques profondes [Snow, 2004], [Zouaq *et al.* 2011] [Zouaq *et al.* 2012], [Mousavi *et al.* 2014].

2.4.1.4 Conclusion

En conclusion, en plus de l'utilisation des techniques statistiques et linguistiques, plusieurs approchent ont eu recours à des techniques d'analyse distributionnelle pour l'automatisation de plusieurs tâches de conceptualisation (*i.e.* extraction de termes, détermination des termes synonymes et l'extraction des relations *est-un* ou ad-hoc). Cependant, les résultats de ces techniques ne sont pas toujours faciles à interpréter. En plus, ils sont parfois erronés à cause des erreurs possibles dans le traitement linguistique de textes.

Certains travaux basés sur la fouille de données ou sur une approche probabiliste utilisent des instances. Cependant, les ontologies à développer n'ont pas toujours beaucoup d'instances [Huang *et al.* 2014]

[18] NN : nom, VB : verbe
[19] ** implique la possibilité de l'existence d'un mot ou d'un ensemble de mots

L'utilisation des ressources externes semble très intéressante mais elle est confrontée à la problématique de la polysémie de termes (problème évident dans WordNet) et sa performance dépend de la qualité de la ressource utilisée.

Enfin, il n'y a pas une seule technique reconnue et privilégiée par tâche du processus de conceptualisation. Et évidement il n'y a pas une technique qui couvre toutes les tâches de ce processus.

La combinaison de plusieurs techniques (*e.g.* apprentissage de patrons, techniques basées sue l'analyse distributionnelle, techniques de raisonnement logique) en utilisant des connaissances profondes nous semble prometteuse [Schropp *et al.* 2013]. Une des perspectives de notre projet de recherche est de développer une approche semi-automatique de conceptualisation d'ontologie combinant plusieurs techniques et utilisant une ontologie noyau pour filtrer et valider les résultats obtenus (cf. le dernier chapitre).

Concernant la conceptualisation d'ontologie (ou plus largement de bases de connaissances à partir du web), la majorité des techniques présentées dans cette section peut être appliquée sur des données du web avec deux difficultés : la sélection et le nettoyage des documents du web et ensuite l'adaptation des techniques utilisées à un nombre important de données [Banko *et al.* 2007], [Wong *et al.* 2012], [Suchanek, 2014], [Mitchell *et al.* 2015].

2.4.2 Cadre pour la comparaison des méthodes et outils de construction d'ontologie

Dans cette section, nous présentons le cadre que nous avons proposé pour comparer des approches semi-automatiques de conceptualisation d'ontologie et leurs outils.

En TAL, des protocoles et métriques ont été proposés pour évaluer des outils d'extraction de termes [Nazarenko *et al.* 2009]. Pour l'analyse des méthodes/outils de conceptualisation automatique des ontologies à partir de textes, trois types de critères ont été proposés dans la littérature [Park *et al.* 2011] : (1) critères concernant des aspects génériques, (2) critères concernant l'extraction et (3) critères concernant la qualité de l'ontologie construite. Nous avons identifié trois limites dans ces travaux [Gherasim, 2013]. La première limite est due à leur hypothèse de travail : la construction automatique d'une ontologie est supposée réalisée à l'aide d'un seul outil, alors que dans la littérature plusieurs approches ont proposé l'utilisation successive de plusieurs outils, correspondant à des tâches différentes, combinée parfois à des tâches manuelles (*e.g.* OntoLearn et Asium).

La deuxième limite concerne les critères portant sur des aspects génériques et l'extraction des artefacts de l'ontologie (*e.g.* le critère « degré d'extraction » doit concerner non seulement les concepts et les relations, mais aussi les instances).

La troisième limite concerne les critères de comparaison de la qualité sémantique et pragmatique des ontologies construites. Certains critères semblent ambigus, difficilement applicables et leur pertinence peut être discutable. Par exemple, la définition proposée pour la consistance est ambigüe puisque usuellement la consistance d'une ontologie fait référence à la consistance logique. De plus, si les critères de précision, complétude et pertinence sont adaptés à l'évaluation de la qualité pragmatique, la méthode proposée pour leur application, basée uniquement sur le jugement humain, est discutable.

Pour répondre à ces limites nous avons proposé dans les travaux de thèse de Gherasim [2013] une méthode de comparaison des approches et de leurs outils en trois étapes. La première étape de comparaison porte sur le degré de complétude et d'automatisation des approches. La deuxième étape de comparaison porte sur les caractéristiques techniques des outils. Elle couvre la plupart des critères génériques et d'extraction proposés par Park. Les critères que nous avons proposés

sont : (1) la disponibilité de l'outil (proposée également par Park) ; (2) le type de ressources que les outils prennent en entrée, la forme des résultats obtenus et le prétraitement des ressources d'entrée pour que les outils puissent les utiliser (extension du critère prétraitement de Park) ; (3) la réutilisation et amélioration (critère différent de celui de Park concernant la possibilité de la réutilisation des résultats de l'outil ou d'autres résultats) ; (4) le degré d'extraction (extension du critère proposé par Park aux instances, aux relations taxonomiques et aux autres relations sémantiques et pas uniquement les concepts et les relations comme dans Park) ; (5) le degré d'automatisation (même critère que celui proposé par Park) ; (6) le paramétrage (extension d'un critère de Park); (7) l'efficacité (même critère que celui proposé par Park) : le temps nécessaire aux outils pour extraire les différents éléments ; (8) la fiabilité (même critère que celui proposé par Park) : on vérifie si les résultats obtenus sont les mêmes lors d'exécutions répétées dans des conditions identiques ; (9) les outils et ressources externes (critère sans correspondance avec ceux de Park) : on vérifie si les outils, en plus des corpus qu'ils prennent en entrée, utilisent, durant le processus d'extraction, d'autres ressources ou des outils externes.

La troisième étape de comparaison porte sur la qualité des résultats. Nous avons proposé une comparaison plus précise que celle proposée par Park. Dans un premier temps les résultats de chaque outil sont comparés à une ontologie construite manuellement. Ces comparaisons permettent une première évaluation quantitative par deux mesures classiques (la précision et le rappel) pour chaque outil. Nous les avons complétées par deux autres mesures : « la spécificité des termes extraits » et « la richesse sémantique des relations taxonomiques ». La spécificité des termes extraits est basée sur la considération suivante : il est plus probable que les termes complexes, composés de deux ou plusieurs mots, correspondent à des concepts plus spécifiques d'un domaine que les termes simples, composés d'un seul mot. La richesse sémantique des relations taxonomiques est basée sur l'observation qu'une relation taxonomique entre concepts dont les étiquettes peuvent s'inclure lexicalement apporte moins d'informations qu'une relation taxonomique entre deux concepts dont les étiquettes n'ont pas cette propriété.

Fig. 2.1 présente une vue globale des relations entre les éléments utilisés dans notre cadre de comparaison des approches et des outils. Ces éléments ont été classés en trois groupes : (1) le groupe *Approche* concerne les éléments permettant la caractérisation d'une approche (les tâches, les techniques utilisées pour l'automatisation, les outils et leur configuration) ; (2) le groupe *Sorties et statistiques* concerne les éléments permettant la caractérisation des résultats obtenus dans les tests (l'ontologie construite automatiquement avec ses composants, et les mesures quantifiant les performances des outils) ; et (3) le groupe Ligne de comparaison concerne les éléments servant de support pour la comparaison des approches (le référentiel de tâches de MethOntology, l'ontologie construite manuellement et le résultat de l'analyse technique des outils).

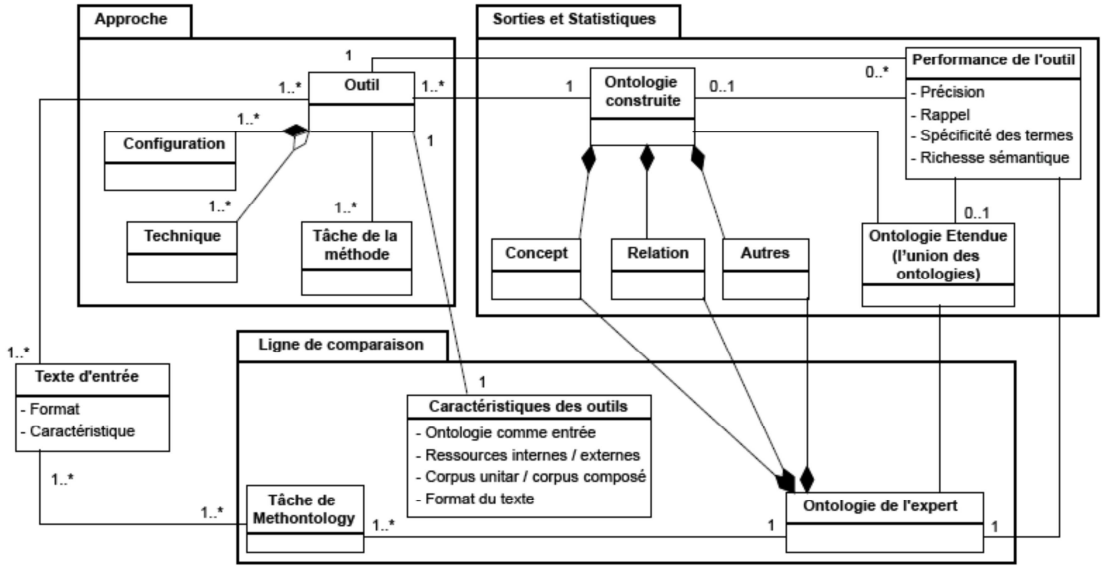

Fig. 2.1- Cadre de comparaison des méthodes et outils de conceptualisation d'ontologie [Gherasim, 2013]

Nous avons appliqué ce cadre pour comparer quatre approches et leurs outils : OntoLearn [Velardi *et al*. 2005], Alvis [Nédellec, 2007], Text2Onto [Cimiano&Volker, 2005], SPRAT [Maynard *et al*. 2009a,b]. D'autres approches/méthodes ont été proposées pour la construction ou l'évolution (semi)automatique d'ontologie associées à des outils logiciels : Terminae [Szulman, 2013], [Szulman, 2011], [Aussenac *et al.* 2008], *Dafoe* [Charlet *et al.* 2010], OntoGen [Fortuna *et al.* 2007], OntoLT [Buitelaar et al. 2004], OntoPlus [Novalija et al. 2011**]**, CA Manager [Damljanovic *et al.* 2009], DYNAMO [Sellami *et al.* 2011], OntoCmaps [Zaouq *et al.* 2012], OntoHarvester [Mousavi *et al.* 2014].

Nous avons comparé seulement ces quatre approchent parce qu'à l'époque (pendant la thèse de T. Gherasim) il n'y avait qu'elles (d'après nos connaissance) qui satisfaisaient les conditions suivantes : (1) elles construisent une ontologie du domaine à partir d'un corpus ; (2) elles automatisent au moins l'extraction des concepts (ou plus correctement des termes) et des relations taxonomiques (ou plus correctement des relations taxonomiques lexicales) ; (3) elles sont génériques et peuvent être potentiellement appliquées à tous les domaines, et (4) elles sont accompagnées d'outils logiciels qui les rendent opérationnelles.

Nous avons donc analysé ces quatre approches en trois étapes, comme mentionné précédemment. Pour la première étape de comparaison, nous avons analysé ces approches et leur(s) outil(s) selon le référentiel de tâches de MethOntology. Cette comparaison nous a permis de mettre en évidence, pour chaque approche : (1) les étapes automatisées et les tâches pour lesquelles l'intervention manuelle des utilisateurs est nécessaire ; (2) les techniques utilisées pour automatiser les tâches ; (3) les éléments de l'ontologie qui peuvent être identifiés. Tab. 2.1 synthétise les principales correspondances entre les tâches de conceptualisation de MethOntology et les tâches qui composent les quatre approches analysées. Il présente également les types de techniques utilisés par chaque approche pour la réalisation de ces tâches. Nous avons indiqué, à l'aide d'une série d'acronymes, la famille de techniques utilisée : S (approches statistiques), L (approches linguistiques), St (techniques structurelles), Di (technique distributionnelle), MP (techniques à base de patrons définis manuellement), PL (techniques à base de patrons appris

pendant le processus de conceptualisation), PA (utilisation de patrons), ES (utilisation de ressources externes) et M (tâche effectuée manuellement).

<table>
<tr><th>MethOntology</th><th colspan="2">Tâches de OntoLearn</th><th>Taches de Alvis</th><th>Tâches de Text2Onto</th><th>Tâches de SPRAT</th></tr>
<tr><td>Construction du glossaire de termes
Identification et définition des termes correspondant aux concepts, attributs, instances et relations.</td><td colspan="2">Extraction des termes – L&S
Filtrage des termes à l’aide de filtres L&S
Validation des termes – M.
Identification de définitions pour les termes à l’aide de recherches sur Internet-ES
Un terme peut correspondre à un concepts</td><td>Extraction des termes–L
Validation des termes – M
Un terme peut correspondre à un concept ou à une instance</td><td>Extraction de termes – L, S
Un terme peut correspondre à un concept ou à une , instance.</td><td rowspan="3">Extraction de termes – L
Un terme peut correspondre à un concept ou à une , instance.

Exécution itérative

Choix d’un nouveau terme impliqué dans un patron est-un le reliant à un concept déjà présent dans l’ontologie – PA

(si le terme choisi est un concept)
Le concept est ajouté à l’ontologie, à l’aide des règles d’intégration et de résolution de conflits (règles spécifiques à SPRAT) – PA

Identification et intégration dans l’ontologie de relations entre le nouveau concept et des concepts déjà présents dans l’ontologie à l’aide des patrons prédéfinis– PA</td></tr>
<tr><td>Construction des taxonomies de concepts
Regroupement des termes synonymes sous le même intitulé
Regroupement des concepts liés par des relations est-un au sein d’une taxonomie</td><td>Désambig. sémantique des termes complexes à l’aide de l’algorithme SSI, spécifique à OntoLearn – ES&St</td><td>Identification de ‘star patterns’ et leur classification– M.
Construction de Word-Class Lattice (WCL) – PL
Identification des correspondances entre les star patterns et le WCL– PA</td><td>Construction d’une taxonomie-
Identification du contexte de chaque terme et construction de la taxonomie par classification non supervisée des termes - Di</td><td>Identification des relations est-un à l’aide de WordNet, des patrons et des heuristiques portant sur les termes composés – St, PA&ES</td></tr>
<tr><td>Construction des diagrammes des relations binaires ad-hoc
Regroupement des verbes synonymes sous le même intitulé
Définition des relations à partir de verbes en rajoutant à chaque verbe des couples de concepts : sujet et complément d’objet</td><td colspan="2">Identification des exemples de relations sémantiques du domaine – M.
Apprentissage de règles pour l’identification des relations– PL.

Utilisation de règles pour identifier des relations entre les composants des termes complexes – PA.</td><td>Identification des dépendances syntaxiques indiquant une relation -M
Identification des relations à l’aide de l’apprentissage inductif basé sur le formalisme ASA-PL
Identification des concepts connectés par les relations apprises à partir d’exemples – PA</td><td>Identification des relations ad-hoc (définies par un verbe) – (à l’aide de l’analyse des syntagmes verbaux et de la fréquence des termes)) – L&S&PA</td></tr>
</table>

Tab. 2.1- Comparaison des tâches et techniques de quatre approches selon MethOntology [Gherasim, 2013]

Cette analyse a montré que les trois activités de conception d'ontologie de MethOntology ont été considérées dans ces quatre approches, avec des techniques diverses. Pour une même approche, ces dernières peuvent être différentes d'une activité à une autre. Ceci confirme l'intérêt d'une de nos perspectives qui consiste à combiner des techniques différentes pour la construction semi-automatique d'ontologie.

Tab. 2.2 synthétise les outils associés à chaque approche. A chacune des quatre approches sont associés un ou plusieurs outils. A Text2Onto est associé un seul outil, qui porte le même patronyme que l'approche et qui construit directement une ontologie à partir de textes. A SPRAT sont associés 4 outils (SPRAT, JAPE, NEBOnE et TermRaider), mais le premier couvre l'ensemble du processus de construction d'une ontologie et n'utilise pas les différents résultats intermédiaires obtenus par les autres outils. A OntoLearn sont associés 5 outils : TermExtractor, GlossExtractor, SSI, WCL System et C4.5. TermExtractor identifie une liste de termes correspondant à des concepts ; GlossExtractor identifie des définitions correspondant à ces termes ; WCL System identifie les relations taxonomiques présentes dans ces définitions ; SSI permet la désambiguïsation sémantique des termes complexes alors que C4.5 permet l'apprentissage inductif de règles pour l'identification de relations sémantiques. A Alvis sont associés YATEA, BioLG, ASIUM, LINKPARSER et Propal.

Ces outils peuvent être synthétiquement classés en deux grandes familles : les outils génériques développés initialement dans un autre cadre, et ceux qui sont spécifiquement dédiés à la construction automatique d'ontologie. Les outils génériques sont essentiellement des outils d'apprentissage inductif de règles (C4.5, Propal), des analyseurs grammaticaux (LINKPARSER, BioLG), des extracteurs de termes (YATEA) et des outils de traitement de patrons textuels (JAPE). Etant donné que ASIUM est le seul outil spécifique proposé par l'approche Alvis et que, à notre connaissance, il n'était pas disponible pour les tests, nous ne l'avons pas considéré dans la troisième étape de comparaison.

Dans cette troisième étape, nous avons réalisé une étude expérimentale préliminaire pour tester ces outils et comparer leurs résultats pour l'extraction des termes et des relations *est-un* entre eux, en utilisant un corpus de 4000 mots portant sur le domaine de construction d'ontologies à partir de textes. Pour l'extraction des termes, nous avons testé TermExtractor, Text2Onto et SPRAT et nous avons comparé leurs résultats par rapport à l'ontologie construite manuellement à partir du même corpus. Cette étape a montré que TermExtractor a la meilleure précision pour l'extraction de concepts, alors que Text2Onto a un meilleur rappel ; SPRAT a un très faible rappel ; Text2Onto extrait plus de termes simples que TermExtractor. Pour l'extraction des relations *est-un*, nous avons comparé les résultats de WCL system et de Text2Onto en utilisant le même corpus. Les résultats ont montré une complémentarité entre les deux outils. Chacun identifie des relations *est-un* de nature spécifique : Text2onto extrait des relations sémantiquement pauvres, en se basant sur une technique structurelle et WCL extrait des relations sémantiquement riches, utilisant des ressources externes. En plus, WCL semble plus performant (il a la même précision que Text2Onto mais un rappel beaucoup plus important). Toutefois, pour la suite de notre travail, nous avons décidé de continuer à utiliser Text2Onto pour une raison d'accessibilité, de gratuité et de possibilité de faire des changements dedans. En effet, au contraire de Text2onto, qui est accessible, gratuit et open, WCL ne peut être utilisé qu'en demandant à ses auteurs de le tester sur nos données.

Pour plus de détails sur ces trois étapes de comparaison des quatre approches OntoLearn, Alvis, Text2Onto et SPRAT, le lecteur peut se référer aux articles [Gherasim et al. 2011a ; 2011b ; 2013b] et à la thèse de T. Gherasim [2013].

Tab. 2.2 - Correspondances entre outils et tâches pour chaque approche

Tâches de MethOntology	Tâches d'OntoLearn				Les tâches d'Alvis		Les tâches de Text2Onto		Les tâches de SPRAT			
	Sous-tâche		Outil		Sous-tâche	Outil	Sous-tâche	Outil	Sous-tâche	Outil		
Construction du glossaire de termes	Extraction des termes Filtrage des termes		*TermExtracto L&S*		Extraction des termes	*YATEA*	Extraction des concepts (termes)	*Text2Onto module Concept L&S&PA*	Extraction des termes	*JAPE NEBONE L*	*Term Raider L&S*	*SPRAT*
	Validation des termes		—		Validation des termes	—	Extraction des instances	*Text2Onto module Instance L&S&PA*	Choix d'un terme		—	
	Identification des définitions pour les termes		*GlossExtracto ES*									
Construction des taxonomies de concepts	Désambig. sémantique des termes complexes	Extraction des – relations *est-un*	*SSI ES &St*	*WCL System PL, PA*	Construction d'une taxonomie	*BioLG, ASIUM adapté pour ASIUM Di*	Identification des relations *est-un*	*Text2Onto module L&S&PA&ES*	Insertion d'un concept dans l'ontologie à l'aide d'une relation *est-un*		—	
	Identification des relations *est-un*		—									
Construction des diagrammes des relations binaires ad-hoc	Identification des exemples de relations sémantiques du domaine		—		Analyse ASA et Choix d'un ensemble d'exemples	*LINK PARSER*	Identification des relations ad-hoc	*Text2Onto module Relation L&PA*	Identification et intégration des relations dans l'ontologie		—	
	Apprentissage de règles		*C4.5 PL*		Apprentissage de règles	*Propal PL* —						
	Identification des relations entre les composants des termes complexes		—		Identification des relations	—						

Dans la section suivante, nous présentons brièvement les améliorations que nous avons apportées à l'outil Text2Onto version anglaise et sa nouvelle version pour la construction d'ontologie à partir de textes en français.

2.4.3 Amélioration de Text2Onto et son adaptation à la langue française

2.4.3.1 Amélioration de la version Text2Onto pour la langue anglaise

Text2Onto est une version enrichie de TextToOnto. Il comprend des algorithmes ou des techniques pour extraire des termes, des instances, des relations *est-un*, des relations ad-hoc et des axiomes de « disjointness[20] ». Text2Onto comprend un modèle POM ((Probalistic Ontology Model) dans lequel une valeur de pertinence est assignée à chaque terme extrait. Au fur et à mesure qu'une ontologie se construit et avec l'ajout de nouveaux textes, chaque POM peut évoluer. Ainsi, on peut retracer l'évolution d'une ontologie suivant l'ajout de documents à un corpus.

Text2Onto fait appel à des librairies externes pour la constrcution d'ontologie. Tout d'abord, il va prendre en entrée un document texte ou un corpus de documents. Il va ensuite faire appel à GATE (General Architecture for Text Engineering) qui est un outil de traitement automatique de la langue naturelle. GATE possède de nombreux plugins dont ANNIE qui est un Part-Of-Speech Tagger pour la langue anglaise. ANNIE va annoter les documents à traiter en déterminant les catégories grammaticales de chaque mot. Une fois les document annotés, des patrons d'extraction des termes et des relations lexicales[21] définis en JAPE (Java Annotation Patterns Engine) vont y être appliqués.

Ces deux dernières étapes s'effectuent grâce à l'outil GATE. Une fois les deux étapes effectuées, le POM (Probalistic Ontology Model) va calculer l'intérêt de chaque terme en fonction de la méthode de calcul choisie par l'utilisateur. Enfin, Text2Onto va établir l'ontologie finale avec les relations et les termes extraits et ensuite sélectionnés par l'utilisateur. Il va produire un fichier .owl correspondant à l'ontologie.

Il est à noter que Text2Onto peut être utilisé en autonomie ou utilisé sur la plateforme NeOnToolkit.

Nous avons testé Text2Onto[22] sur plusieurs corpus de textes et nous avons identifié plusieurs problèmes dans ses résultats [Gherasim *et al.* 2012 ; 2013a ; 2013b] [Chulyadyo&Mittal, 2012] [Benard&Vourch, 2014]. Plusieurs points faibles ont été identifiés, certains ont été traités, voire supprimés [Chulyadyo, 2012]. Tout d'abord, le manque de richesse de ses résultats (*e.g.* des termes extraits souvent composés d'un seul mot, des relations *est-un* avec un apport sémantique faible[23], peu de relations ad-hoc identifiées). Ceci est dû au fait que Text2Onto utilise principalement une technique structurelle et des patrons lexico-syntaxiques pour identifier des termes, des relations ou des axiomes de « disjointness ». En plus, le nombre de patrons lexico-syntaxiques utilisés est très réduit. Nous avons commencé par rajouter des nouveaux patrons, que ce soit pour l'identification des termes, des relations *est-un* ou ad-hoc. Le test de la version améliorée de Text2Onto sur l'article [Navigli&Velardi, 2004] a donné une augmentation très

[20] Dans ce manuscrit, nous utilisons ce terme en anglais parce que nous n'avons pas trouvé un consensus pour sa traduction

[21] Il s'agit bien sûr de termes ou de relations terminologiques puisque le passage de termes à concepts n'est pas considéré dans Text2Onto

[22] Des étudiants du master EM-DMKM et des étudiants ingénieurs en Informatique Décisionnelle de polytech-Nantes ont participé au test de Text2Onto.

[23] Dans nos travaux, nous considérons qu'une relation *est-un* entre deux termes struturellement reliés (ontologie noyau *est-un* ontologie) a un apport sémantique moins important que celui d'une relation est-un entre deux termes qui ne le sont pas [Gherasim *et al*. 2013b].

considérable du nombre de relations extraites (de plus de 300%) [Chulyadyo&Mittal, 2012], [Chulyadyo, 2012], [Morineau *et al.* 2013]. D'autres problèmes ont été identifiés et analysés (cf. chapitre suivant). Cependant, le problème majeur dans Text2Onto est la méthode d'extraction des relations ad-hoc. En effet, dans Text2Onto, des patrons permettent d'extraire des relations ad-hoc entre termes. Si une relation ad-hoc est extraite plusieurs fois mais avec des domaines différents (de même pour des co-domaines différents), Text2Onto fait le choix de choisir comme domaine (de même pour le co-domaine) le terme le plus fréquent dans le corpus traité, ce qui peut donner des résultats aberrants [Chulyadyo, 2012], [Benard&Vourch, 2014]. Cependant, nous n'avons pas traité ce problème.

2.4.3.2 Adaptation de l'outil Text2Onto à la langue française.

Dans le projet KIFANLO, pour pouvoir traiter des textes en français, nous avons adapté Text2Onto à la langue française. Un des points forts de Text2Onto est la possibilité de l'adapter à une langue en interchangant des composants de traitement linguistique ou des algorithmes. D'ailleurs, text2Onto a été déjà adapté à l'espagnol et à l'allemand [Volker *et al.* 2008].

Pour ce faire, nous avons tout d'abord remplacé l'outil de POS tagger Annie pour l'anglais par le POS tagger TreeTagger pour le français. Nous avons remplacé les patrons en anglais par des patrons existants pour la langue française que nous avons définis en JAPE [Aussenac&Jacques, 2008].

Par ailleurs, dans le projet KIFANLO, nous avons remarqué la nécessité d'adapter les patrons d'extraction des relations ad-hoc au domaine de l'ontologie à construire et aux textes traités. Nous avons donc rajouté dans Text2Onto adapté à la langue française d'autres patrons liés aux relations noyaux de l'ontologie à développer et à leurs synonymes (*i.e.* synonymes des concepts et des relations noyaux). Le test de la première version française sur des textes dans le domaine de la pêche a montré des résultats insuffisants mais encourageants pour continuer à l'améliorer et lui intégrer d'autres techniques pour la construction semi-automatique d'ontologie, en utilisant une ontologie noyau[24] [Benard&Vourch, 2014], [Benard *et al.* 2014].

2.5 Notre expérience en construction d'ontologie

Dans chacun des trois projets UEML[25], ISTA3 et KIFANLO, nous avons participé à la construction d'une ou de plusieurs ontologies.

UEMO. Dans le projet UEML**,** nous avons développé l'ontologie UEMO (Unified Enterprise Modelling Ontology) [Opdahl *et al.* 2012]. UEMO porte sur le domaine des langages de modélisation d'entreprise. Son objectif est d'aider à faire interopérer des applications hétérogènes en définissant des correspondances entre les langages de leurs modèles. En effet, UEMO est utilisée pour annoter des constructs des langages de modélisation. Les résultats des annotations sont comparés à l'aide d'une mesure sémantique afin de déterminer des similarités ou correspondances sémantiques entre les constructs. UEMO et son ontologie noyau ont été définies en empruntant quand c'est nécessaire, des concepts et relations de l'ontologie de haut niveau BWW [Bunge, 1977], [Wand&Weber, 1988]. L'ontologie noyau a été construite selon la structure du méta-modèle UEML (cf. section 4.2 du chapitre 4). Ce dernier est utilisé comme support pour aider à annoter les constructs d'un langage. L'ontologie noyau d'UEMO comprend 4 concepts noyaux : Class, Property, State et Transformation et des relations noyaux entre eux. Chaque concept noyau a été spécialisé en plusieurs concepts pour former un module ou une

[24] Cf. Section 3.3.1.4
[25] Il s'agit du groupe de travail UEML du Rex Interop

taxonomie. Le module Class (Fig. 2.2) comprend 37 concepts, le module Propriety (Fig. 2.3) comprend 51 concepts, le module Transformation comprend 6 concepts et le module State comprend 9 concepts. UEMO a été formalisée en logique de description [Opdhal *et al.* 2012]. Elle est aussi définie en Protege-OWL. Plusieurs plugins pour protege-OWL, utilisant cette ontologie ont été développés [Anaya *et al.* 2010] :

- un plugin pour annoter les éléments d'un langage de modélisation d'entreprise avec UEMO ;
- un plugin pour vérifier la cohérence des annotations réalisées et la définition de l'ontologie UEMO ;
- un plugin pour mesurer la similarité sémantique des éléments de langages annotés avec UEMO.

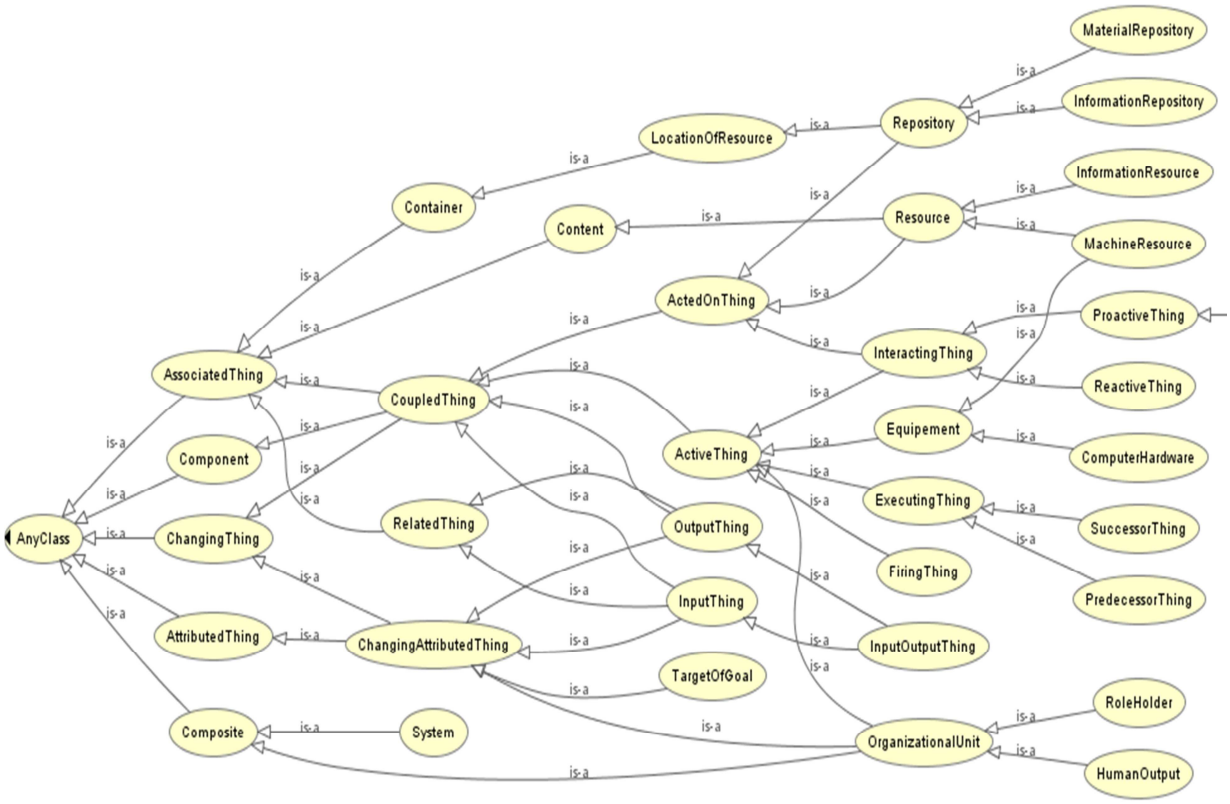

Fig. 2.2 - Extrait de la taxonomie Class d'UEMO

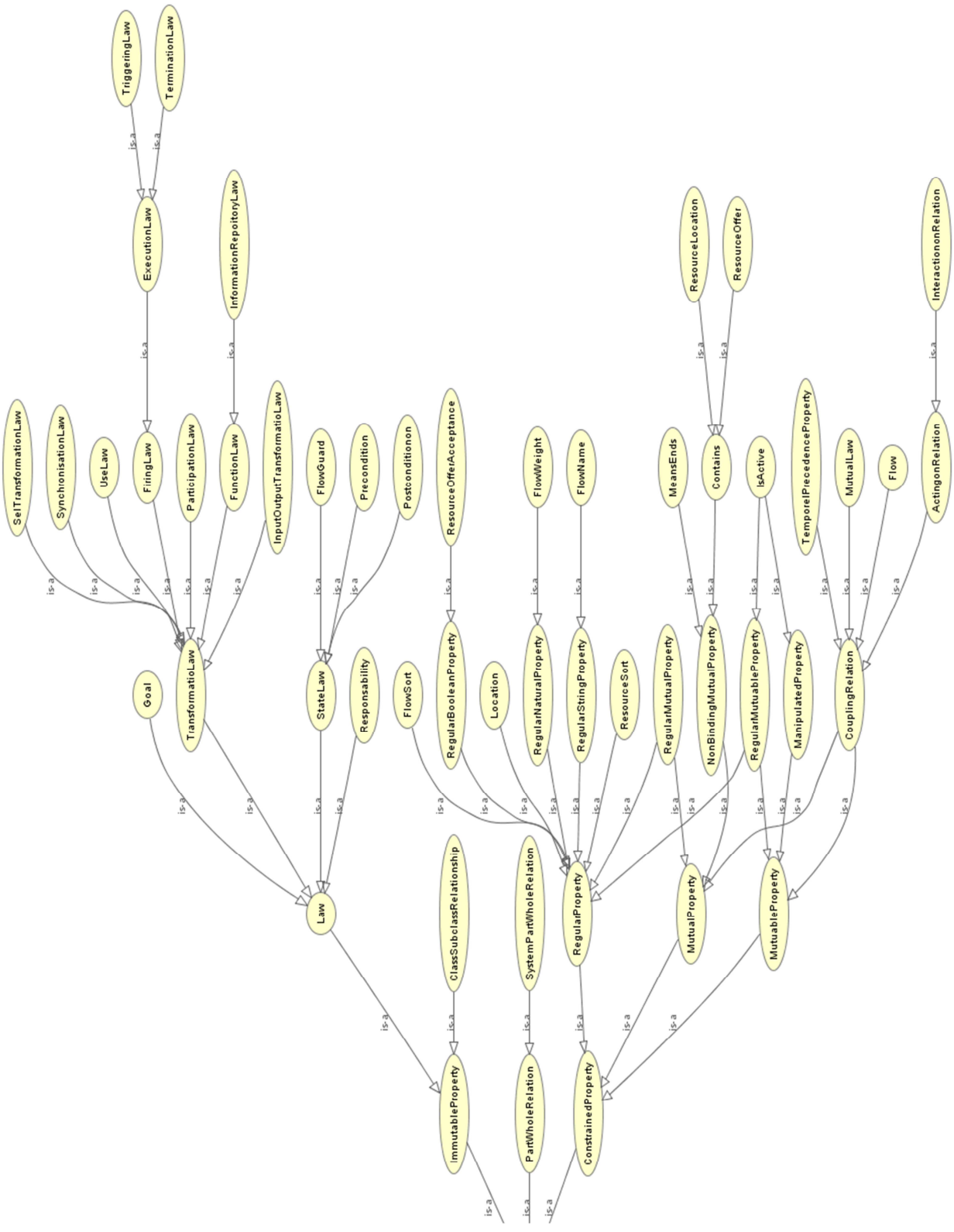

Fig. 2.3 - Extrait de la taxonomie Property d'UEMO

ISTA3. L'objectif des ontologies développées dans le projet FUI ISTA3 est d'aider à faire interopérer des applications hétérogènes de plusieurs sous-traitants de l'aéronautique. Dans ce projet, nous avons tout d'abord défini une architecture hybride pour l'interopérabilité des systèmes : une ontologie globale et une ontologie locale pour chaque partenaire industriel [Harzallah, 2012], [Harzallah *et al.* 2014]. Ces ontologies ont porté en particulier sur le domaine de la conception, la fabrication et la vente des pièces en composite. Nous avons défini en premier lieu une ontologie noyau de ce domaine comprenant cinq concepts noyaux : Produit, Processus, Activité, Ressource et Elément conceptuel. Nous l'avons enrichie en utilisant l'AP239 (PLCS) de la norme ISO 10303 (STEP) et des concepts extraits de l'ontologie développée dans [Dartigues, 2003] portant sur la géométrie de produit**,** sans considérer le domaine de pièces en composite. Nous avons voulu ensuite développer l'ontologie de conception/fabrication/vente des pièces en composite. Cependant, sa construction manuelle nécessitait une expertise métier forte dont l'acquisition s'est heurtée à la faible disponibilité des experts. Nous nous sommes orientés donc vers une construction semi-automatique à partir de corpus de textes techniques. Pour maîtriser la complexité de la construction de cette ontologie, nous avons procédé par développement modulaire. Nous avons considéré l'ontologie noyau spécialisée à un certain niveau pour pourvoir choisir avec les partenaires du projet les concepts à spécialiser chacun en un module (Fig. 2.4). Six modules ont été retenus : M1-Elément conceptuel, M2-Matière composite, M3-Processus d'usinage de Composite, M4-Processus de gestion de devis, M5-Processus de gestion de facture et M6-Processus de gestion de commande. Nous avons ensuite, déterminé pour chacun les textes à traiter, dont la pertinence a été jugée par les experts métiers. Ces textes sont extraits de trois types de documents : (1) Documents Wikipédia (7000 mots), (2) Glossaire (9000 mots) et (3) Aide de l'ERP SAGE X3 (40000 mots). Nous avons utilisé Text2Onto pour construire chaque module. Les liens entre chaque module et l'ontologie noyau ont été déterminés à la main.

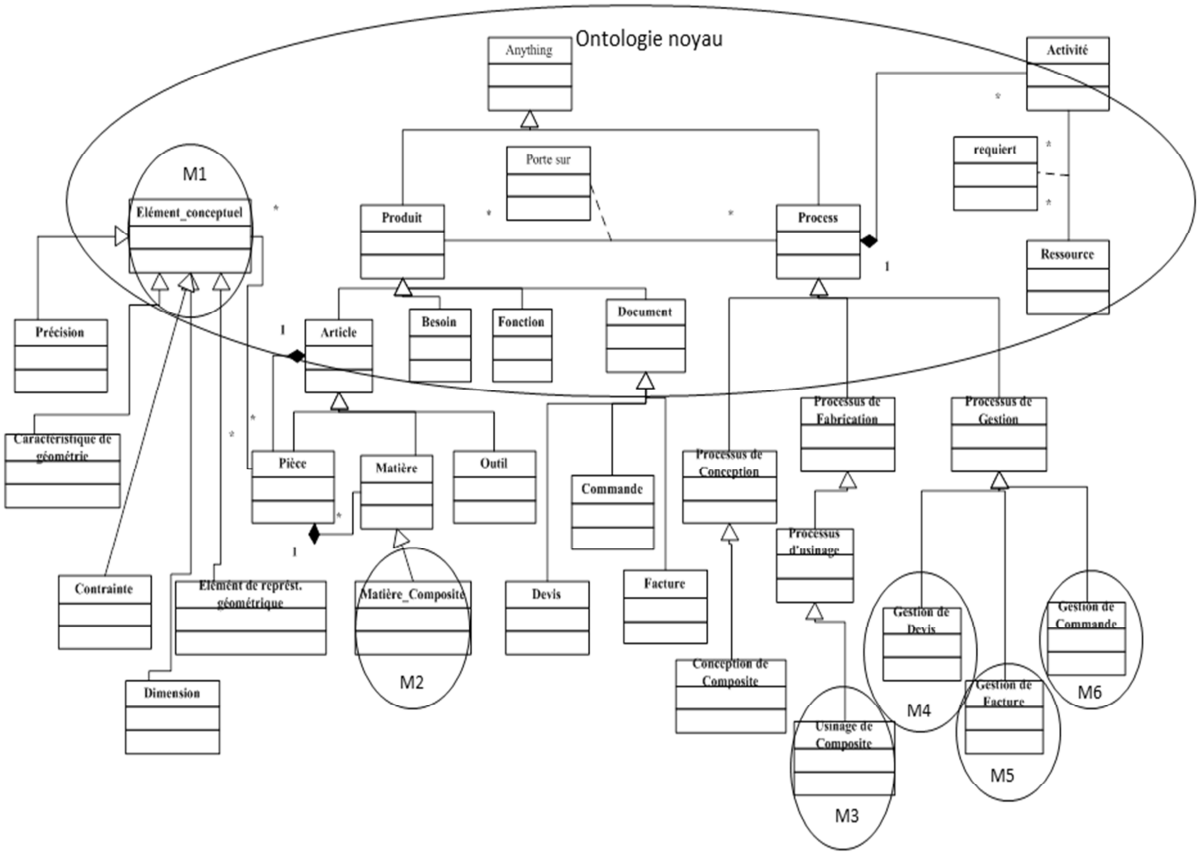

Fig. 2.4 - Premier niveau de spécialisation de l'ontologie noyau de ISTA3

KIFANLO. Le projet KIFANLO s'intéresse aux pratiques de la pêche, leur dynamique spatiale et leur évolution dans le temps. La capitalisation de ces pratiques a été réalisée en partie grâce à des enquêtes auprès de 100 pêcheurs. Cependant la diversité des vocabulaires des pêcheurs et l'interdisciplinarité du domaine de la pêche, a dirigé le projet vers le développement de l'ontologie KIFANLO pour représenter, structurer, formaliser, mutualiser les connaissances des pêcheurs et comparer leurs stratégies de pêche. En effet, après une ou deux interviews avec un pêcheur, sa stratégie de pêche a été représentée par une carte cognitive dont les nœuds sont des concepts de l'ontologie KIFANLO. Pour concevoir l'ontologie KIFANLO, nous avons adopté une approche semi-automatique basée sur une ontologie noyau. Les étapes réalisées sont : (1) la définition d'une ontologie noyau et le choix de la modélisation de ses composants, (2) l'enrichissement de cette ontologie par des taxonomies existantes, (3) l'enrichissement semi-automatique et incrémental de cette ontologie (choix des textes et utilisation de Text2onto), et (4) le remaniement de l'ontologie noyau, en ciblant mieux l'objectif de l'ontologie développée. Actuellement, l'ontologie KIFANLO comprend 410 concepts avec 7 concepts noyaux (Fig. 2.5).

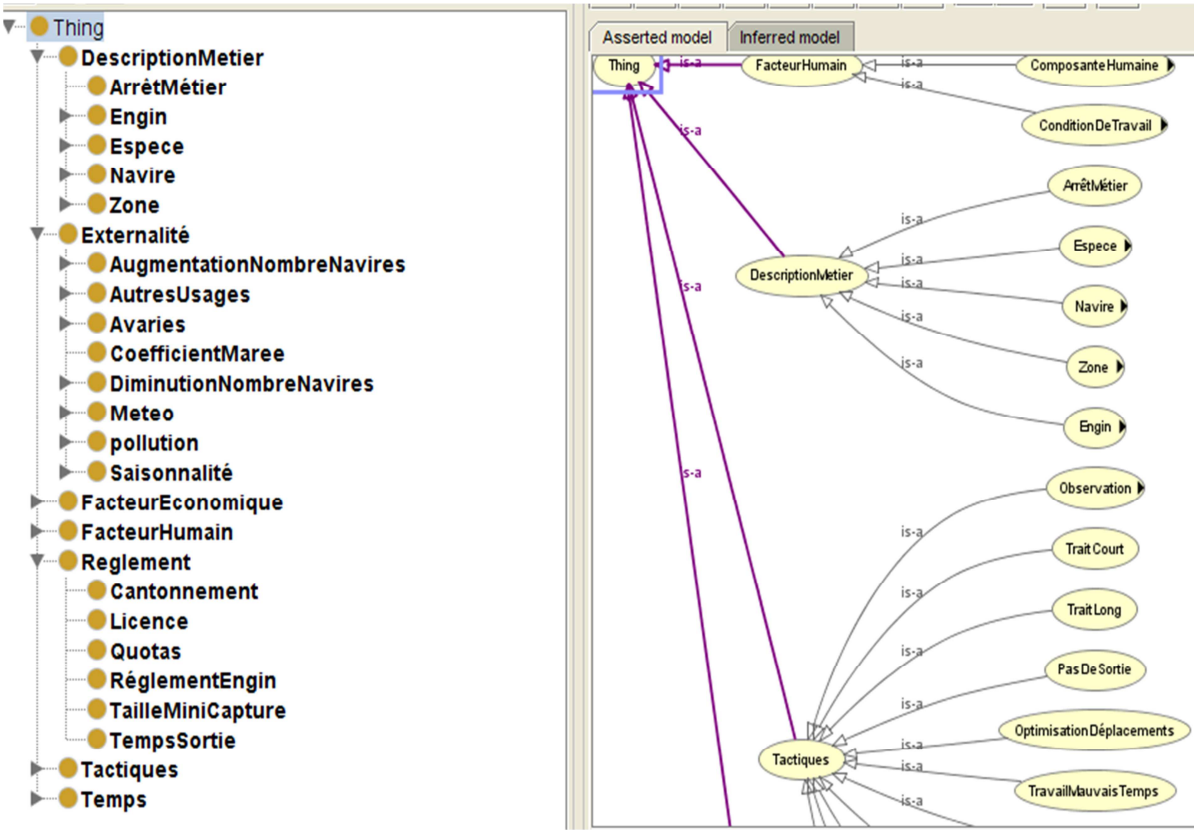

Fig. 2.5 - Extrait de l'ontologie KIFANLO

Approche adoptée et difficultés rencontrées.

Dans chacun de ces trois projets, nous avons adopté une approche de conceptualisation basée sur une ontologie noyau, tout en suivant l'approche MethOntology. Les deux intérêts principaux de l'utilisation d'une ontologie noyau dans ces projets sont :

- de gérer la complexité de la construction d'une ontologie. C'était le cas pour la construction d'UEMO. Cette dernière porte sur un domaine abstrait et elle a été définie d'une façon collaborative par plusieurs experts du domaine[26]. Pour gérer la complexité de sa construction, nous[27] avons tout d'abord commencé par nous mettre d'accord sur la sémantique des concepts de son ontologie noyau et des relations entre eux et les formaliser. Ensuite, nous avons défini les concepts d'UEMO par module. Chaque module correspond à une taxonomie ayant un concept noyau pour racine.
- de bien délimiter le domaine d'une ontologie notamment parce que l'ontologie à construire porte sur un domaine vaste (c'était le cas des ontologies de KIFANLO et ISTA3)

Nous n'avons pas rencontré de problèmes pour la définition d'une ontologie noyau dans UEML et ISTA3. L'ontologie noyau d'UEMO a été définie à partir d'une ontologie existante de haut niveau. L'ontologie noyau dans ISTA3 est composée de concepts bien maitrisés par les

[26] Les experts étaient Giuseppe Berio université de Bretagne Sud , Andreas Opdahl et Raimundas Matulevičius université de Bergen , Victor Anaya et Maria Jose Verdecho université de Valencia, Patrick Heymans université de Namur , Michele Dassisti université de Bari, Herve Panetto université de Nancy, et moi-même.
[27] Les experts du domaine.

membres du projet et représentant bien le domaine de l'ontologie à développer. Cependant, nous n'avons considéré que certains de ces concepts noyaux afin de limiter la taille de l'ontologie ISTA3.

Dans les trois projets, chaque concept noyau a été spécialisé en sous-concepts en discutant avec des experts, en cherchant des parties d'ontologies existantes qui le détaillent ou en identifiant des textes portant sur ce concept. Ces textes ont été ensuite traités à la main ou d'une façon semi-automatique pour en extraire des concepts ou des relations que nous avons ensuite liés à la main aux concepts noyaux.

De plus, dans les projets UEML et KIFANLO, les ontologies ont été construites pour annoter des objets afin de les comparer sémantiquement. Nous avons donc utilisé une ontologie noyau comme un moyen pour structurer et simplifier le processus d'annotation (cf. section 3.3.2). La formalisation de l'ontologie noyau d'UEMO a demandé beaucoup de temps afin d'aboutir à un consensus entre les différents experts du domaine. Mais, cette ontologie a permis de bien structurer par la suite la phase de développement d'UEMO. Le manque de formalisation de l'ontologie noyau de KIFANLO a engendré des confusions dans la sémantique des concepts noyaux. Après de multiples modifications, nous avons eu recours à la définition de chaque concept noyau en langue naturelle afin de régler en partie ce problème de confusion sémantique et aider à bien classifier des concepts sous les concepts noyaux.

Dans les projets KIFANLO et ISTA3, nous avons considéré la construction semi-automatique à partir de textes. Pour gérer la complexité de ce type de construction, nous avons développé chaque concept noyau comme un module de l'ontologie à développer, construit d'une façon semi-automatique à partir d'un ensemble de textes qui lui sont spécifiques et validé séparément. En plus dans le projet KIFANLO, nous avons enrichi l'ontologie avec des concepts identifiés dans les interviews des pêcheurs sur les stratégies de pêche.

Dans les projets UEML et KIFANLO les différents membres du projet étaient très motivés pour contribuer au développement de l'ontologie, parce qu'ils voyaient bien son objectif et son rôle dans le projet. Le développement a été réalisé relativement vite et les améliorations ont été demandées, voire proposées par les utilisateurs du projet. Dans le projet ISTA3, tous les membres du projet n'étaient pas convaincus de l'intérêt d'une ontologie. Pour eux, il y avait très peu de vocabulaires différents d'un industriel à un autre et ils arrivaient à se comprendre sans ontologie. Cette situation a freiné le processus de développement d'ontologie.

Dans le projet UEML, une des difficultés rencontrées est la complexité de la sémantique des concepts d'UEMO. En effet, il s'agit de concepts de haut niveau, abstraits et difficiles à expliciter leur sémantique. Evidemment, la formalisation a aidé les différents experts du domaine à mieux se comprendre et à se mettre d'accord sur leur sens.

Dans le projet KIFANLO, nous avons été confrontés au choix du contenu d'une ontologie noyau (les concepts noyaux et leur niveau d'agrégation) et de sa modélisation (concept ou attribut, attribut ou relation…). Plusieurs remaniements de cette ontologie ont été réalisés. Nous avons essayé de faire face à ce problème en rappelant chaque fois l'objectif de l'ontologie à développer.

La phase de conception automatique d'ontologie avec Text2onto n'a pas été facile à réaliser pour plusieurs raisons. Tout d'abord, il y avait le problème du choix des textes et de la façon de les traiter (*e.g.* tous ensemble ou par étape). Ensuite, les résultats obtenus n'étaient pas vraiment intéressants au début. A cette étape, il était nécessaire de bien comprendre les techniques utilisées dans Text2onto et les adapter aux ontologies à développer. Ensuite, il y avait le problème du filtrage et de la validation des résultats. Pour les deux ontologies, nous avons rencontré le problème de traitement de la langue française. Dans ISTA3, nous avons décidé de construire une

ontologie bilingue : en français et en anglais. En effet, nous avons traité des textes en anglais et nous avons associé à chaque label en anglais un label en français. Dans le projet KIFANLO, nous avons décidé d'adapter Text2Onto à la langue française.

Enfin, dans ISTA3 et KIFANLO, nous avons utilisé aussi l'ontologie noyau pour améliorer la classification des termes extraits d'une façon semi-automatique sous les concepts noyaux. Dans ces deux projets, cette tâche a été réalisée à la main et elle s'est avérée relativement fastidieuse. Par ailleurs, nous avons développé une méthode pour améliorer la classification de termes par rapport aux concepts noyaux, mais nous n'avons pas eu l'occasion de l'appliquer dans ces deux projets. Cette méthode est présentée dans le chapitre suivant, section 3.3.1.4.

2.6 Conclusion

Dans ce chapitre, nous avons présenté l'évolution des approches de construction d'ontologie avec l'évolution des ressources à partir desquelles des ontologies sont construites, des techniques et outils pour leur construction ainsi que de l'objectif de leur exploitation.

Nous nous sommes focalisés particulièrement sur l'étape de conceptualisation. Nous avons remarqué le manque de méthode pour rendre opérationnelle la conceptualisation de grandes ontologies pour un cas réel. En outre, nos expériences en construction d'ontologie nous ont montré la pertinence d'une approche modulaire basée sur une ontologie noyau formelle, pour la conceptualisation de ce type d'ontologie.

L'accessibilité des données a créé un engouement pour la construction semi-automatique des ontologies à partir de textes. Dans ce chapitre, nous avons présenté différentes techniques proposées dans la littérature pour automatiser l'extraction de termes, de relations de subsumption ou de relations ad-hoc à partir de textes. Les techniques basées sur une analyse distributionnelle sont très intéressantes car elles permettent de proposer des classifications de termes ou de relations. Cependant, dans la majorité des cas, il faut bien revoir ces classes pour définir leur sémantique et celles des relations entre elles. Les techniques utilisant les patrons peuvent s'appliquer sur n'importe quel type de texte. Leur rappel reste faible parce qu'il faut les adapter au domaine et au style des textes étudiés. L'analyse profonde de textes semble très importante pour l'apprentissage de patrons afin de faire face à la complexité grammaticale d'un texte. Enfin, des nouvelles techniques restent à développer pour l'étape d'extraction des axiomes : peu de travaux traitent cette question [Volker *et al.* 2007], [Petasis *et al.* 2013], [Alec, 2016], [Haidar-Ahmad *et al.* 2016].

Pour aider à choisir une approche ou un outil de construction semi-automatique d'ontologie, nous avons proposé un cadre de comparaison suivant trois étapes : la comparaison du degré de complétude et de l'automatisation des approches, la comparaison des caractéristiques techniques des outils et la comparaison des résultats des outils. Notre cadre complète les études existantes en remédiant à leurs points faibles.

La comparaison d'outils pour la construction semi-automatique d'ontologie nous a révélé le manque d'outils couvrant toutes les tâches de la conceptualisation. Par ailleurs, ces outils sont intéressants lors de la première phase de traitement automatique (et donc rapide) de textes pour l'extraction de termes et de relations terminologiques. Cependant, la qualité de leur résultat est insuffisante et demande une phase antérieure d'adaptation de l'outil au domaine de l'ontologie à développer et une phase postérieure importante de filtrage et de correction de l'ontologie obtenue. L'amélioration des résultats d'un outil ou leur correction requièrent aussi une compréhension voire une maîtrise des techniques utilisées dans cet outil, ce qui n'est pas une tâche facile si ce dernier n'est pas bien documenté ou n'est pas fidèle à sa description.

L’intégration de plusieurs techniques et outils existants pour la construction d’ontologie pourrait être pertinente pour couvrir plusieurs tâches de la conceptualisation et se servir des résultats obtenus par ces différents outils à partir d’un même corpus, comme un moyen de validation.

Enfin, à partir de nos expériences en construction d’ontologie, nous avons identifié le rôle important d’une ontologie noyau formelle pour guider la conceptualisation d’ontologie. Nous présentons dans le chapitre 5 des perspectives qui portent sur la construction semi-automatique d’ontologie basée ontologie noyau, en combinant des techniques différentes.

Chapitre 3 : Vers une approche semi-automatique de construction et validation intégrées d'ontologie

Sommaire

3.1 Introduction

La validation d'ontologie est une étape incontournable dans le processus de construction d'ontologie. Sans elle, l'ontologie ne pourrait pas être exploitable. Cette étape devient de plus en plus complexe avec la taille croissante des ontologies et le recours à leur construction semi-automatique.. En effet, l'accès facile à des masses de données dans des bases documentaires ou sur le web a motivé le développement d'outils pour l'automatisation de la construction d'ontologie. Toutefois, comme nous avons vu dans le chapitre précédent, leur utilisation dans des projets réels ont montré l'insuffisance de la qualité de leurs résultats.

Généralement, la validation d'ontologie peut être considérée selon trois critères principaux : (1) les *dimensions de l'ontologie à évaluer* (*e.g.* dimension fonctionnelle, dimension structurelle ou dimension d'usage) [Gangemi *et al.* 2006], [Duque-Ramos *et al.* 2011] ; (2) la méthode de validation (manuel vs automatique) [Vrandecic, 2009] et le profil de l'utilisateur, s'il y en a (*e.g.* ingénieur de connaissances, utilisateur, expert du domaine de l'ontologie) [Hartmann, 2004], (3) et l'étape à laquelle elle est effectuée (*e.g.* pendant le processus de son développement, avant sa publication) [Hartmann, 2004], [Tartir *et al.* 2010].

Des approches ont été proposées pour analyser la qualité d'une ontologie et la valider. Actuellement, il n'y a pas encore de consensus sur la manière dont une ontologie doit être validée [Neuthaus&Vizedom, 2013]. Des outils ont été développés pour identifier certains problèmes logiques (*e.g.* l'inconsistance logique ou l'insatisfiabilité) dans une ontologie et aider à les supprimer. Malheureusement, certains autres problèmes (*e.g.* l'inconsistance sémantique) ne sont pas encore considérés et sont identifiables que par l'humain, quand la taille de l'ontologie le permet. L'humain est souvent sollicité pour la validation d'ontologie, mais dans la majorité des cas, il n'intervient qu'à la fin de son processus semi-automatique de conceptualisation. Certains travaux ont recommandé une meilleure intégration de l'humain dans ce processus de conceptualisation [Simperl&Tempich, 2009] et plus particulièrement de bien intégrer l'étape de validation dans ce processus et le rôle de l'humain dans cette étape.

C'est pour cela que nous prônons un processus de conceptualisation composé de deux processus menés en parallèle et en coopération : (1) le processus d'extraction/acquisition et (2) le processus de validation. Le premier processus se focalise sur l'acquisition ou l'extraction des termes pertinents pour le domaine de l'ontologie à construire et les liens entre eux et l'identification et le nommage de ces connaissances pertinentes (telles que les concepts et les rôles) alors que le deuxième est un processus garantissant l'obtention d'une ontologie avec la qualité attendue. Le processus de validation devrait être réalisé le plus tôt possible afin d'éviter la propagation des problèmes et rendre complexe leur correction. En outre, nous considérons la validation comme un processus (1) qui cherche une mauvaise qualité dans une ontologie : il s'agit du sous-processus d'«évaluation de la qualité d'une ontologie» ; et (2) qui propose ensuite une méthode pour améliorer cette qualité en identifiant et en enlevant les causes liées à cette mauvaise qualité (*i.e.* modification/suppression/rajout d'artefacts à l'ontologie en cours de construction) : il s'agit du sous-processus de « correction d'une ontologie ».

Dans la littérature, l'évaluation de la qualité d'une ontologie est appréhendée selon trois facettes : (1) le score associé à la qualité d'une ontologie (tel que « élevé », « moyen », « faible »… ou des valeurs numériques) déterminé par une ou plusieurs mesures de qualité (selon Gangemi *et al.* [2006]) ou donné par un expert ou par une application qui utilise cette ontologie, (2) les problèmes de qualité *i.e.* symptômes des défauts ou défauts potentiels qui impactent la qualité d'une ontologie, et (3) les défauts dans une ontologie *i.e.* défauts portants sur ses artefacts et qui sont la cause de cette mauvaise qualité/problème [Harzallah *et al.* 2015]. Ces trois facettes

sont évidemment reliées. Par exemple, une ontologie inconsistante du point de vue logique est un problème et c'est le symptôme d'un défaut dans cette ontologie ; les défauts sont les axiomes qui engendrent cette inconsistance ; un score de qualité peut être défini en fonction du nombre d'axiomes qui causent cette inconsistance (une mesure de la qualité). Cependant, même s'il semble évident que les mesures de qualité soient liées aux défauts, ce n'est pas toujours le cas. Par exemple, Gangemi *et al.* [2005] considèrent que la mesure de profondeur d'une ontologie entre dans l'évaluation de « l'ergonomie cognitive d'une ontologie », une des dimensions de la qualité d'une ontologie, sans faire référence à aucun défaut potentiel lié à cette mesure. De même un score de qualité donné par un expert n'est pas nécessairement accompagné par la liste des problèmes et des défauts à corriger dans une ontologie. Ou encore, les résultats erronés donnés par une application utilisant une ontologie pourraient nous renseigner sur une mauvaise qualité de cette ontologie, mais ils ne nous indiquent pas les défauts à corriger.

L'utilisation en pratique d'un score de qualité pour valider une ontologie n'est pas facile, même si elle pourrait quand-même nous avertir d'une mauvaise qualité en l'absence d'un processus complet de validation. Par contre, les problèmes introduits comme des symptômes des défauts semblent plus efficaces pour valider (*i.e.* enlever les défauts éventuels dedans) une ontologie que l'utilisation d'un score de la qualité d'ontologie.

Dans ce chapitre, nous présentons et discutons des méthodes d'évaluation et de validation d'ontologie (section 3.2). Nous montrons la variabilité des types de problèmes pouvant nuire à la qualité d'une ontologie et le manque de standardisation de la terminologie utilisée pour cette notion de problème. Dans la section 3.3, nous présentons nos contributions pour le développement de deux approches de conception et validation intégrées d'ontologie : une première approche basée sur les problèmes et une deuxième approche par l'annotation d'objets, et bien sûr, appliquée aux ontologies pour l'annotation d'objets.

Dans le cadre de la première approche, nous avons proposé quatre contributions. La première est une typologie évolutive des problèmes qui englobe les types de problèmes traités dans la littérature et qui les synthétise d'une façon uniforme (section 3.3.1.1). A partir des liens déterminés entre des problèmes de notre typologie, nous avons défini un ordre pour identifier des problèmes dans une ontologie en parallèle du processus de sa conceptualisation, permettant d'éviter la propagation de problèmes[28] dans une ontologie (section 3.3.1.2). Comme troisième contribution, nous avons proposé des anti-patrons partiels et une heuristique pour identifier le problème de « Contradiction sociale » dans une ontologie, tout en minimisant le nombre de questions à poser à un acteur social [Harzallah, 2016] (section 3.3.1.3). Ces anti-patrons sont basés sur des liens formels entre des problèmes de « Insatisfiabilité » et des problèmes de « Contradiction sociale ». Notre quatrième contribution (section 3.3.1.4) est une méthode pour aider à la correction du problème « Ontologie plate » utilisant des règles inductives définies à partir d'une ontologie noyau. Son objectif est de mieux structurer une ontologie en identifiant des nouvelles relations *est-un* entre les termes/concepts extraits et les concepts noyaux d'une ontologie.

Notre deuxième approche de construction et de validation intégrées d'ontologie s'intègre dans le cadre de l'évaluation d'une ontologie par son utilisation dans une application (section 3.3.2). Elle porte sur une ontologie formelle et utilise des règles et des guidelines pour détecter qu'une mauvaise qualité lors de son utilisation pour l'annotation d'un objet et/ou lors de son enrichissement.

[28] *i.e.* quand un problème provoque d'autres problèmes dans une ontologie et rend complexe leur identification.

En conclusion, nous articulons nos différentes contributions pour définir une méthodologie semi-automatique de construction et validation intégrées d'ontologie.

3.2 Méthodes d'évaluation d'ontologie

Comment nous l'avons expliqué dans l'introduction, l'évaluation de la qualité d'une ontologie peut être appréhendée selon trois facettes : (1) le score associé à la qualité de cette ontologie déterminé par une ou plusieurs mesures de qualité ou donné par un expert ou par une application qui utilise cette ontologie, (2) les problèmes de qualité dans cette ontologie, et (3) les défauts dans cette ontologie. Fig. 3.1 représente ces trois facettes (*i.e.* mesures de qualité, problèmes de qualité et défauts) et des relations clés entre elles. Des techniques peuvent exister ou sont à développer pour détecter des problèmes de qualité bien définis (a). Des mesures de qualité peuvent être utilisées pour la détection des défauts ou problèmes dans une ontologie, surtout quand elles sont reliées à des valeurs de référence et à des problèmes ou défauts spécifiques (b). Les mesures de qualité sont aussi des techniques pour évaluer la qualité (c). Les problèmes dans une ontologie peuvent être utilisés comme un moyen qualitatif pour évaluer sa qualité (d). Quand un problème de qualité est avéré on peut utiliser une ou plusieurs techniques pour le corriger (e). Enfin, il est possible qu'à certains problèmes, il n'y ait pas de technique associée.

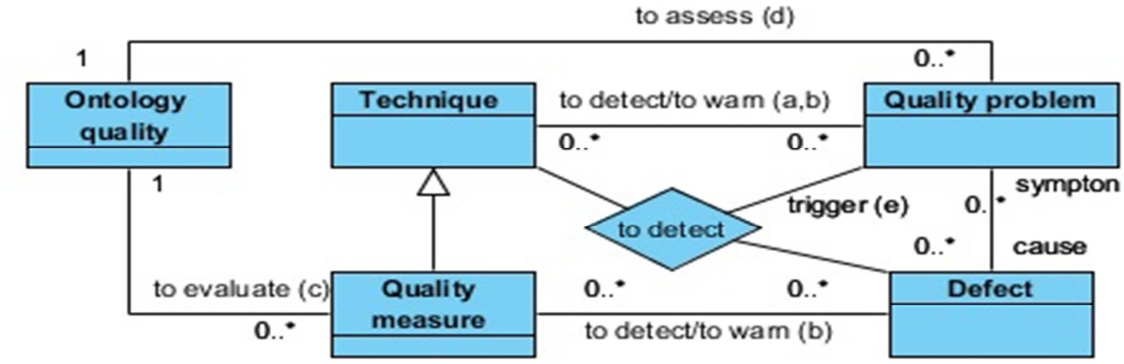

Fig. 3.1 - Les trois facettes de l'évaluation de la qualité d'une ontologie et les relations clés entre elles ([Harzallah *et al.* 2015])

Mesures de qualité. Plusieurs mesures de qualité ont été proposées couvrant plusieurs dimensions de la qualité d'ontologie. La proposition la plus complète associant dimensions de l'ontologie et mesures de qualité est probablement oQual [Gangemi *et al.* 2006]. Dans la proposition oQual, une ontologie est analysée selon trois dimensions : (1) dimension structurelle (la syntaxe et la sémantique formelle d'une ontologie), (2) dimension fonctionnelle (la sémantique intentionnelle[29] d'une ontologie et de ses composants) et (3) dimension d'usage portant notamment sur la facilité de la compréhension des différents artefacts de l'ontologie. Une liste de mesures est associée à chaque dimension pour évaluer sa qualité. Par exemple, les mesures de profondeur et de largeur ont été proposées pour la dimension structurelle ; les mesures de précision et de rappel du contenu d'une ontologie par rapport à ce qu'elle doit représenter ont été proposées pour la dimension fonctionnelle ; et le nombre d'annotations dans une ontologie a été proposé pour la dimension d'usage. En général, bien que l'intérêt des mesures

[29] The intended meaning.

proposées ait été bien expliqué dans la littérature, certaines d'entre elles sont toujours difficiles à lier aux défauts et aux problèmes d'une ontologie. Par exemple, le rapport entre le nombre de concepts et le nombre de relations (NbConcepts/NbRelations) est une mesure de qualité pour évaluer « l'ergonomie cognitive » d'une ontologie qui est elle-même très liée à la « facilité d'utilisation » d'une ontologie. Cependant, une ontologie peut avoir un faible score pour cette mesure sans pour autant présenter de problèmes ou de défauts liés à ses artefacts. C'est pourquoi, la définition des mesures de qualité, comme une entrée à l'évaluation de la qualité, ne fait pas le lien d'une façon explicite aux défauts ou aux problèmes potentiels dans une ontologie. Par contre, considérer les problèmes pour évaluer une ontologie permet de mettre en évidence ses défauts et ensuite de les corriger.

Problèmes de qualité. Plusieurs termes sont utilisés dans la littérature pour évoquer la notion de problème affectant la qualité d'une ontologie : (1) les erreurs de taxonomie [Gomez-Perez *et al.* 2001] ou les erreurs structurelles [Buhmann *et al.* 2011], (2) les anomalies de conception ou défaillances [Baumeister&Seipel, 2005], (3) les anti-patrons [Roussey *et al.* 2010], [Buhmann *et al.* 2011], (4) les embûches « pitfalls » ou les mauvaises pratiques [Poveda *et al.* 2009], [Roussey *et al.* 2010], [Buhmann *et al.* 2011].

Les erreurs de taxonomie sont de trois types : l'inconsistance, l'incomplétude et la redondance [Gomez-Perez *et al.* 2001]. Trois classes d'inconsistance logique ou sémantique ont été considérées : les erreurs de circularité (*e.g.* un concept qui est une spécialisation de lui-même), les erreurs de partitionnement (*e.g.* un concept qui est une spécialisation de deux concepts disjoints), et les erreurs sémantiques (*e.g.* une relation de spécialisation en contradiction avec les connaissances de l'utilisateur). L'incomplétude correspond, par exemple, à des relations ou des axiomes manquants. Finalement, la redondance existe quand, par exemple une relation de spécialisation peut être déduite par des inférences logiques.

Les anomalies de conception concernent la compréhension et la maintenabilité d'une ontologie [Baumeister&Seipel, 2005 ; 2010] : les « lazy concepts » (des concepts feuilles n'ayant pas d'instance et qui ne sont impliqués dans aucune relation ou axiome de l'ontologie en question), la chaine d'héritage (une longue chaine de spécialisation avec un fils unique), et les « lonely disjoint concepts » (des axiomes de « disjointness[30] » superflus qui portent sur des concepts très éloignés dans une ontologie).

Les anti-patrons sont des expressions connues qui engendrent des problèmes identifiables [Roussey *et al.* 2009], [Buhmann *et al.* 2011]. Certaines classes de ces anti-patrons sont : des anti-patrons logiques (produisant des conflits qui peuvent être détectables par un raisonnement logique), des patrons cognitifs (causés par une mauvaise compréhension des conséquences logiques des axiomes), des guidelines (des expressions complexes correctes du point de vue logique et du point de vue cognitif, mais pour lesquelles des alternatives simples ou plus précises existent).

Les embûches (Pitfalls) couvrent des problèmes pour lesquels des patrons de conception d'ontologie « ontology design patterns » (ODPs) ne sont pas disponibles (cf. chapitre 2).

Des classifications de la notion de problème ont été déjà proposées. Gomez-Perez *et al.* [2001] ont défini une typologie de trois classes portant sur les erreurs de taxonomies : les inconsistances, les incomplétudes et les redondances.

Poveda *et al* [2012] ont classé les embûches en sept classes, en fonction de l'axe de qualité qu'elles dégradent. Ces sept classes sont-elles mêmes positionnées par rapport aux dimensions de l'ontologie, à savoir la dimension structurelle, la dimension fonctionnelle et la dimension

[30] Dans ce manuscrit, nous utilisons ce terme en anglais parce que nous n'avons pas trouvé un consensus pour sa traduction.

d'usage. Poveda et al [2012] ont essayé aussi d'aligner ces sept classes à la classification de Gomez-Perez *et al.*, mais ce n'était pas faisable pour toutes ces classes.

Buhmann *et al.* [2011] ont proposé une classification permettant de gérer les erreurs dans un sens très large, puisqu'elle contient des classes couvrant non seulement les erreurs commises ou rencontrées lors de la conceptualisation, mais également des fautes de syntaxe (par rapport au langage manipulé) et des problèmes de transfert de fichiers lors de l'utilisation d'une ontologie. Cette classification oppose les problèmes syntaxiques, sémantiques et structuraux aux problèmes de performance du raisonnement et ceux portant sur les données liées. Les problèmes syntaxiques regroupent les problèmes relatifs au langage utilisé. Les problèmes sémantiques sont ceux qui peuvent être détectés par un raisonneur. La catégorie des problèmes structuraux porte, comme son nom l'indique, sur des problèmes de structure dans une ontologie. La catégorie des problèmes de performance du raisonnement comprend des erreurs de conception qui ralentissent grandement le raisonnement. Les problèmes sur les données liées pourront apparaître lors de la récupération de toute ou une partie d'une ontologie (e.g. mauvaise interprétation du format, fichier inexistant). La classification de Buhmann *et al.* [2011] est relativement complète. Elle ne se limite pas aux problèmes relatifs à une ontologie elle-même, mais on y ajoute des problèmes pouvant intervenir à plus bas niveau (langage utilisé) ou à plus haut niveau (connexion d'une ontologie sur le web). Toutefois, certains problèmes existants dans l'état de l'art ne peuvent pas être positionnés dans cette classification, par exemple le problème de polysémie des labels des concepts.

3.3 Nos contributions pour la construction d'ontologie intégrée à sa validation

3.3.1 Approche de validation d'ontologie par les problèmes

Nous avons défini une nouvelle typologie permettant de standardiser la définition des problèmes et aider à les identifier [Gherasim *et al.* 2012], [Gherasim, 2013a] (section 3.3.1.1). Il est important de disposer d'une classification afin de pouvoir, lorsque l'on trouve un nouveau problème, le rapprocher et le comparer à d'autres problèmes qui lui ressemblent, s'assurer que ce n'est pas un problème connu, et éventuellement s'inspirer ou réutiliser des méthodes de détection des problèmes appartenant à la même classe pour l'identifier.

Nous avons déterminé un ordre selon lequel il est préférable d'identifier les problèmes pouvant nuire à la qualité d'une ontologie (section 3.3.1.2). Cet ordre permet de faciliter l'identification de ces problèmes, voire d'éviter l'occurrence de certains d'entre eux. Par exemple, l'identification d'une redondance impliquant deux relations *est-un* peut nous amener à la suppression d'une de ces deux relations. Ensuite, nous pouvons nous rendre compte que celle qui n'a pas été supprimée n'est pas correcte (problème de type S1, cf. Tab. 3.1) et nous la supprimons également. Par conséquent, l'ontologie devient incomplète (rajout d'un problème de types L3 ou S4, cf. Tab. 3.1). Dans cet exemple, la redondance (problème de type L12, cf. Tab. 3.1) est causée par un problème de type S1 et son existence pourrait nous aider à identifier ce dernier.

En plus, nous avons développé une méthode d'aide à l'identification des problèmes de « Contradiction sociale » (problèmes de type S1, cf. Tab. 3.1) en utilisant des anti-patrons partiels (section 3.3.1.3), et une méthode pour la correction du problème de « Ontologie plate » (problème de type S9, cf. Tab. 3.1) en utilisant des règles inductives définies à partir d'une ontologie noyau (section 3.3.1.4).

3.3.1.1 Typologie des problèmes dans une ontologie

Notre typologie considère une ontologie comme un objet sémiotique : il doit être compris et manipulé par une machine, mais aussi compris, interprété et exploité par un acteur social. Elle est inspirée du framework de qualité des modèles conceptuels SEQUAL [Krogstie *et al.* 1995]. Ce cadre est lui-même sémiotique car il distingue bien un modèle comme un artefact formel qui doit être compris et exploité par une machine, du fait qu'il est un artefact interprété et exploité par un acteur social. La notion d'acteur social est une terminologie empruntée à ce framework et implique un individu ou un groupe d'individus, partie prenante du modèle en question.

Notre typologie est définie selon deux dimensions complémentaires : le niveau logique versus le niveau social et le niveau erreur versus le niveau situation indésirable [Gherasim, 2013a], [Harzallah *et al.* 2014] (cf. Tab. 3.1). Les erreurs sont les problèmes qui rendent une ontologie inutilisable et les situations indésirables sont les problèmes qui affectent une ontologie sans la rendre inutilisable. Les problèmes sociaux sont reliés à la perception qu'un acteur social a d'une ontologie et à l'utilisation qu'il fait de cette ontologie. La perception et l'utilisation envisagées peuvent ne pas être exprimées d'une façon formelle. D'une certaine façon, une autre distinction entre la facette sociale et la facette logique d'une ontologie repose sur la différence entre les connaissances tacites et les connaissances explicites. Les problèmes logiques sont soient des problèmes internes à l'ensemble de formules qui composent l'ontologie (contradictions internes, redondances, *etc.*), soient des problèmes dus à des différences entre l'ontologie et ses modèles intentionnels[31]. Les problèmes logiques du premier type sont identifiables d'une façon automatique et l'identification des problèmes du second type nécessite la disponibilité des modèles intentionnels.

Nous avons distingué cinq erreurs logiques : Inconsistance logique, Ontologie inadaptée aux modèles intentionnels, Ontologie incomplète, Inférences incorrectes par rapport aux conséquences logiques, et Inférences incomplètes par rapport aux conséquences logiques.

Les situations indésirables logiques ont un impact négatif sur la qualité non fonctionnelle d'une ontologie, à savoir sa réutilisabilité, sa maintenabilité, son efficacité, *etc.* Nous avons distingué six situations indésirables logiques : Equivalence logique d'artefacts distincts, Artefacts d'ontologie indifférentiables du point de vue logique, Artefacts_OU, Artefacts_ET, Insatisfiabilité, Complexité élevée des inférences, et Ontologie non minimale.

Nous avons distingué quatre types d'erreurs sociales : Contradiction sociale, Perception d'erreurs de conception, Absence de sens du point de vue social, et Incomplétude du point de vue social. Les situations indésirables sociales sont, pour la plupart, liées aux difficultés rencontrées par un acteur social quand il essaie de comprendre et d'utiliser une ontologie. Comme dans le cas des situations indésirables logiques, il est difficile de dresser une liste exhaustive. Nous avons distingué sept situations indésirables sociales parmi les plus communes : Absence ou mauvaise qualité des explications textuelle, Artefacts potentiellement équivalents, Artefacts indifférenciables du point de vue social, Artefacts avec des étiquettes polysémiques, Ontologie plate ou absence de structuration, Ontologie formalisée à l'aide d'un formalisme non standardisé, Absence d'une version certifiée d'une ontologie pour tous les langages standardisés, et Artefacts inutiles dans une ontologie

Tab. 3.1 regroupe ces différents problèmes et propose une formalisation pour certains d'entre eux. Nous avons formalisé la majorité des problèmes logiques en logiques de description en considérant les notions synthétisées dans Guarino *et al.* [2009] *i.e.* Interprétation (I), Modèles

[31] Un modèle intentionnel (Intended Models) est un modèle formel qui respecte l'engagement ontologique d'une conceptualisation [Garino et al. 2009].

intentionnels (IM), Langage (L), Ontologie (O), et les 2 relations : ⊨ et ⊢ [Gherasim, 2013]. Nous avons ensuite formalisé certains problèmes sociaux, en rajoutant la notion de SAP_O (Social Actor Perception of Ontology) par analogie avec la notion de IM. Par exemple, $SAP_O \nvDash \varphi$ implique φ est fausse dans la perception d'un acteur social de O [Harzallah, 2016].

Par ailleurs, dans cette typologie nous avons mis en évidence une analogie ou une correspondance entre les problèmes logiques et les problèmes sociaux pouvant permettre de passer des uns vers les autres. Par exemple, si on identifie un problème de « Contradiction sociale », nous pouvons construire des modèles intentionnels liés à ce problème (ou des contre exemples) et identifier peut-être un problème de « Ontologie non adaptée ». Nous montrons dans la section 3.3.1.2, l'existence de plusieurs types de relations entre problèmes et comment les exploiter pour améliorer le processus d'identification de problèmes.

Enfin, pour évaluer cette typologie, nous avons étudié les problèmes traités dans la littérature et nous avons réussi à les classifier dans notre typologie [Gherasim, 2013], [Cavy, 2015]. Nous avons classifié 60 cas de problèmes traités dans la littérature (cf. Annexe). Nous les avons numérotés de P1 à P60 : 33 cas (de P1 à P33) traités dans le catalogue de Poveda *et al.* [2012], 11 cas (de P34 à P44) considérés dans [Roussey *et al.* 2010], 10 cas (de P45 à P54) proposés dans [Wang *et al.* 2005] et présentés dans [Buhmann *et al.* 2011], 6 cas (de P55 à P60) proposés dans [Fahad *et al.* 2008]. Cette classification a été effectuée en se basant sur la formalisation de chaque problème en logique de description (quand c'est possible). En effet, un problème est affecté à une classe de problème si sa formalisation implique celle de cette classe. Quand la formalisation d'un problème ou d'un cas de problème n'était pas possible, nous avons effectué la classification en utilisant la définition ou l'explication de ce problème. Par exemple :

- Le problème P45.Partition_error est classé en L10 parce que sa formalisation : $\{O \models C \sqsubseteq (C_i \sqcap C_j), C_i \sqcap C_j \equiv \bot\}$ implique celle de L10 (cf. Tab. 3.1) ;
- Le problème P14.Misusing_« Owl : all values-from » est classé en S1 car sa formalisation est $\{O \models C \sqsubseteq \forall R.C_i , SAP_O \Phi C \sqsubseteq \forall R.C_i , SAP_O \sqsubseteq \exists R.C_i \}$ implique (ou un cas particulier de) celle de S1 (cf. Tab.3.1).

Pour plus d'exemples sur la formalisation des problèmes, le lecteur peut se référer à la section 3.3.1.3 où nous utilisons la formalisation des problèmes de type « Contradiction sociale » ou de type « Insatisfiabilité » pour déterminer des liens entre eux.

Problèmes Logiques	
Erreurs	L1. Inconsistance logique : Il n'existe pas une interprétation I telle que I $\models$ O
	L2. Ontologie inadaptée: il existe une formule φ telle que pour certains IM de L, φ est fausse et O $\models$ φ
	L3. Ontologie Incomplète : Il existe une formule φ pour chaque IM de L, φ est vraie et O $\nvdash$ φ
	L4. Inférences Incorrectes : quand une formule φ fausse dans les IM de L et O $\nvdash$ φ, peut être inférée par un système de raisonnement (O $\vdash$ φ)
	L5. Inférence Incomplète: quand une formule φ vraie dans les IM de L et O $\models$ φ ne peut pas être inférée par un système de raisonnement (O $\nvdash$ φ)
Situations Indésirables	L6. Equivalence logique d'artefacts distincts de l'ontologie : O $\models$ $A_i \equiv A_j$
	L7. Artefacts d'ontologie non différentiables du point de vue logique : c'est impossible de montrer que : (O $\models$ $A_i \equiv A_j$), (O $\models$ $A_i \sqcap A_j \sqsubseteq \bot$) ou (O $\models$ {c}$\sqsubseteq A_i$ et {c} $\sqsubseteq A_j$)
	L8. Artefacts d'ontologie « OU »: A_i équivalent à $A_j \cup A_k$, $A_i \neq A_j$, $A_i \neq A_k$, mais il n'existe ni un rôle R (si applicable) tel que (O $\models$ $(A_j \sqcup A_k) \sqsubseteq \exists R.\top$), ni une instance c s.t. (O $\models$ {c} $\sqsubseteq A_j$ et O $\models$ {c} $\sqsubseteq A_k$)
	L9. Artefacts d'ontologie « ET »: A_i équivalent $A_j \sqcap A_k$, $A_i \neq A_j$, $A_i \neq A_k$, mais il n'existe pas un rôle R (si applicable) tel que (O $\models$ $(A_j \sqcup A_k) \sqsubseteq \exists R.\top$)
	L10. Insatisfiabilité: soit un artefact A, O $\models$ $A \sqsubseteq \bot$)
	L11. Complexité élevée des inférences : un raisonnement complexe dans O alors qu'un raisonnement plus simple est possible
	L12. Ontologie non minimale : des informations non nécessaires
Problèmes sociaux	
Erreurs	S1. Contradiction Sociale: contradiction entre l'interprétation que l'acteur social donne à l'ontologie et les axiomes de l'ontologie : O $\models$ φ, $SAP_O \nvdash$ φ (φ est fausse dans la perception de l'acteur social)
	S2. Perception d'une erreur de conception : e.g. modélisation des instances comme des concepts
	S3. Absence de sens du point de vue social : interprétation impossible
	S4. Incomplétude du point de vue social : $SAP_O \models$ φ et O $\nvdash$ φ
Situations Indésirables	S5. Absence ou mauvaise qualité des explications textuelles
	S6. Eléments de l'ontologie potentiellement équivalents : $SAP_O \models A_i \equiv A_j$ et impossible de montrer que O $\models$ $A_i \equiv A_j$
	S7. Eléments d'ontologie non différentiable du point de vue social : impossible de montrer que : ($SAP_O \models A_i \equiv A_j$), ($SAP_O \models A_i \sqcap A_j \sqsubseteq \bot$) ou ($SAP_O \models$ {c}$\sqsubseteq A_i$ et {c} $\sqsubseteq A_j$)
	S8. Eléments d'ontologie avec des étiquettes polysémiques
	S9. Ontologie plate
	S10. Ontologie formalisée avec un langage non standardisé
	S11. Absence de version certifiée de l'ontologie
	S12. Eléments inutiles dans l'ontologie

Tab. 3.1 - Typologie des problèmes ([Gherasim, 2013] et [Harzallah, 2016] avec modifications)

Cette classification est originale parce qu'elle est basée sur la formalisation des problèmes en logiques de description. Elle a montré que la majorité des problèmes traités dans la littérature correspond à un des 5 types de problèmes suivants : « Insatisfiabilité » (12 cas), « Non minimale » (12 cas), « Contradiction sociale » (11 cas), « Erreurs de conception » (12 cas), et « Incomplétude d'un point de vue social » (9 cas). Elle nous a permis d'identifier des liens de généralisation ou d'équivalence entre problèmes [Cavy, 2015]. Par exemple, les problèmes P34 et P53 sont équivalents et les problèmes P47 et P48 le sont aussi (cf. section 3.3.1.3).

3.3.1.2 Ordre d'identification des problèmes

Nous nous sommes intéressés à optimiser l'intervention humaine dans le processus de validation d'une ontologie en optimisant (1) le nombre de problèmes à identifier ou à corriger et (2) le niveau de difficulté d'identification ou de correction de ces problèmes.

Dans un processus de validation, l'ordre suivant lequel on identifie et/ou on corrige des problèmes pourrait optimiser l'intervention humaine dans ce processus. Par exemple :

1. . il n'a pas de sens d'identifier des problèmes de « Ontologie non adaptée » dans une ontologie, s'il y a un problème de « Inconsistance logique » dedans qui n'a pas encore été corrigé. Ceci veut dire qu'il vaut mieux tout d'abord corriger les problèmes d' « inconsistance logique » (qui peuvent être normalement identifiés d'une façon automatique) avant de chercher à identifier des problèmes de «« Ontologie non adaptée » ;
2. vérifier s'il y a une redondance dans une ontologie et enlever ce problème sans avoir préalablement vérifié les problèmes de « Contradiction sociale » ou de « Ontologie non adaptée » dedans peut engendrer un problème de « Ontologie Incomplète ». Ceci veut dire qu'il ne faut pas corriger les problèmes de redondance avant de corriger les problèmes d'« Ontologie inadaptée » ou « Contradiction sociale (cf. section 3.3).
3. Utiliser les équivalences de concepts distincts (problème L6) aide à identifier et corriger les problèmes de « Ontologie inadaptée » ou de « Contradiction sociale ». Cependant, comme précédemment, il ne faut pas corriger L6 avant de corriger les problèmes de « Ontologie inadaptée » ou de « Contradiction sociale ».

Nous avons identifié des liens, nommés « dépendance de validation », entre les problèmes d'une ontologie. Tab.3.2 comprend une liste de dépendances de validation qui contraint l'ordre pour l'identification et la correction de problèmes. Une dépendance de validation « *A1,..,An → B1,…,Bm* » implique qu'avant de vérifier un des Bi, tous les problèmes de Ai (*i.e.* A1 à An) demandent normalement d'être identifiés et supprimés (ou corrigés) par des techniques appropriées.

En pratique, dans un processus de validation, les dépendances peuvent être utilisées selon une approche « avant » (*i.e. forward)* ou une approche « arrière » *(i.e. backward). Par exemple*, pour cette dépendance « *A1,..,An → B1,…,Bm »,* A1,.., An peuvent être vérifiés et corrigés avant de considérer B1,…,Bm (Approche Avant) ou B1,…, Bm peuvent être vérifiés afin d'aider à cibler A1,…,Am qui doivent être ensuite corrigés avant de corriger B1,…, Bm (Approche Arrière).

Il est à noter que supprimer un problème logique d'une ontologie implique l'obtention d'une nouvelle ontologie qui n'est pas nécessairement équivalente à la précédente d'un point de vue logique. Il est clair que supprimer un problème de type : L1, L2, L3 ou L10 peut se faire mais sans une équivalence logique avec l'ontologie précédente alors que des problèmes de type : L4, L5, L6 ou L12 peuvent être supprimés en garantissant des équivalences logiques avec l'ontologie précédente. La suppression des problèmes de type social engendre normalement une nouvelle ontologie non équivalente à la précédente.

Ces dépendances de validation ont été identifiées en faisant une analyse informelle de leur signification. L'origine de certaines dépendances provient de l'étude que nous avons réalisée des problèmes identifiés dans deux ontologies construites d'une façon semi-automatique avec Text2Onto à partir de deux textes de natures différentes. Le premier texte est un article scientifique de 4500 mots portant sur la construction d'ontologie à partir de texte et le deuxième texte est un glossaire technique de 9500 mots portant sur le domaine des produits en composite. Nous avons identifié la cause de chaque problème et les problèmes qu'il aurait pu causer. Nous avons également analysé les corrélations des occurrences des problèmes dans les deux ontologies [Gherasim *et al.* 2012], [Harzallah *et al.* 2014]. Ces corrélations nous ont guidé dans l'identification de certaines dépendances de validation.

Ces dépendances ne sont pas encore validées. Nous sommes en train d'étendre ces dépendances de validation à des nouvelles dépendances, de les formaliser et de les prouver, en se

basant sur la formalisation de chaque problème. Enfin, nous sommes également en train de développer une démarche pour définir un ordre d'identification et de correction des problèmes, basé sur ces dépendances de validation. Cet ordre pourrait être adapté aux résultats de chaque étape de conceptualisation, en fonction des types de problèmes qui pourraient exister dedans, afin d'intégrer la validation à la conceptualisation. Ces travaux en cours représentent une des perspectives de notre projet recherche (cf. Chapitre 5).

L1 → L2, L3, L4, L5, L6, L7, L8, L9, L10, L11, L12	Si une ontologie est inconsistance d'un point de vue logique, les inférences n'ont pas de sens. Cette dépendance ne peut être utilisée qu'en dépendance « avant ».
L2, L3 → L6, L7, L8, L9, L10, L11, L12 L2, L3 → L4, L5	Il n'a pas de sens d'identifier des situations indésirables dans une ontologie qui n'a pas été complétement finalisée ; la même chose est vraie pour un raisonnement incorrect ou incomplet. Cette dépendance doit être utilisée en dépendance « avant » ; mais son utilisation en dépendance « arrière » est possible (par exemple, L6 peut être identifié et supprimé et ceci peut être utilisé pour mettre en évidence les IM à définir)
S12, S3 → S1, S2, S5, S6, S7, S8	Les artefacts inutiles et incompréhensibles doivent être enlevés avant de vérifier et identifier d'autres problèmes. S10 et S11 peuvent être identifiés indépendamment.
S1, S2, S3, S12 → L2,L3,L4,L5,L6,L7,L8,L9,L10,L11, L12	Les artefacts inutiles et incompréhensibles, les erreurs de conception et de contradiction sociale doivent être vérifiées avant les problèmes logiques (à l'exception de L1) ; Cette dépendance peut être aussi utilisée en dépendance « arrière ». Par exemple, L12 est vérifié, les artefacts redondants sont identifiés et des problèmes S1 portant sur ces artefacts redondants sont vérifiés et si nécessaires supprimés.
S2 → S1,S4,S9	Erreurs de conception doivent être supprimées avant de vérifier les problèmes de Contradiction sociale, d'Incomplétude ou d'Ontologie Plate. Cette dépendance a été utilisée en dépendance « «avant ».

Tab. 3.2 - Dépendances de validation entre les problèmes de qualité (traduit de [Harzallah *et al.* 2015])

3.3.1.3 Anti-patrons partiels pour l'aide à l'identification de « Contradiction sociale »

Plusieurs travaux se sont intéressés aux techniques d'identification ou de correction des problèmes pouvant affecter la qualité d'une ontologie. Des raisonneurs (Pellet, FactC++, Racer) ont été développés pour identifier certains problèmes logiques, particulièrement les problèmes de « Inconsistance logique » (L1), de « Insatisfiabilité » (L10) ou de redondance (faisant partie du problème « Ontologie non minimale » (L12)), mais sans déterminer leur origine. D'autres travaux ont cherché à identifier et à présenter à un acteur social le plus petit ensemble d'axiomes menant à une insatisfiabilité [Wang *et al.* 2005], [Rodler *et al.* 2013]. Pour déterminer la cause ayant la meilleure chance d'être la bonne et en minimisant le nombre de questions à poser à un acteur social, Rodler *et al.* [2013] utilisent une méthode probabiliste prenant en compte les causes les plus fréquentes d'un problème. Des heuristiques ont été proposées pour détecter des problèmes qui ne peuvent pas être identifiés par des raisonneurs, par exemple des candidates à l'absence d'axiomes de « disjointness » [Quazi&Quadir, 2011]. Roussey *et al.* [2010] ont défini des anti-patrons pour l'identification de certains problèmes de « Insatisfiabilité » (L10) ou des candidats à des problèmes d'« Ontologie inadaptée » (L2) ou d'« Ontologie non minimale » (L12).

L'identification des problèmes de « Ontologie inadaptée » (L2) et de « Incomplétude d'un point de vue logique » (L3) demande des IM, qui ne sont pas toujours évidents à avoir. L'identification des problèmes de type L6, L7, L8 ou L9 peut se réaliser à l'aide de techniques simples.

Concernant les problèmes sociaux, certaines approches utilisent des systèmes de question/réponse pour les identifier [Pammer, 2010]. Les questions portent souvent sur la totalité des artefacts d'une ontologie. Ceci permet d'aboutir à une validation complète mais qui repose sur un acteur social qui pourrait commettre des erreurs en répondant à des centaines de questions. Poveda *et al.* [2012] ont développé l'outil OOPS ! pour aider à détecter certains de ces problèmes. Enfin, la méthode LOVMI (Les Ontologies Validées par Méthode Interactive) combine des approches et outils (*e.g.* le raisonneur Hermit, OOPS !, validation manuelle) pour identifier et corriger certains problèmes logiques ou sociaux dans une ontologie d'une façon interactive et collaborative [Richard *et al.* 2015].

En conclusion, la majorité des travaux s'est intéressé à l'identification et à la correction des problèmes d'insatisfiabilité en faisant intervenir ou non un acteur social. En revanche, peu de travaux se sont intéressés à la détection des problèmes sociaux.

Pour aider à identifier les problèmes de type « Contradiction sociale », nous nous sommes basés sur la formalisation des liens entre ce type de problème et d'autres problèmes. Il s'agit pour deux problèmes P1 et P2 et leur formule respective F(P1) et F(P2) de déterminer quelle information F nécessaire pour que F(P1) & F soit équivalente à F(P2). Comme nous avons expliqué précédemment, des liens existent entre certains problèmes. D'une façon intuitive, des liens existent entre les problèmes d'« Insatisfiabilité » et les problèmes d'« Inconsistance logique », « Ontologie inadaptée » ou de « Contradiction sociale ». En effet, un problème d'« Insatisfiabilité » représente un ou plusieurs artefacts dont l'interprétation est vide. Cependant, on définit rarement ce type d'artefacts. Ceci implique que probablement, on possède des instances pour l'artefact insatisfiable et en les rajoutant à l'ontologie on produit un problème de « Inconsistance logique ». Ce dernier cache souvent l'existence de quelque chose qui est faux dans cette ontologie par rapport à des IM (problème de « Ontologie inadaptée ») ou par rapport à la perception d'un acteur social (problème de « Contradiction sociale »).

Par exemple, le problème P36 de « Insatisfiabilité » (cf. Tab.3.3) dans une ontologie O où F(P36)= **{**$O \models Ci \sqsubseteq \forall R.Cj$, $Ci \sqsubseteq \exists R.C_k$, $C_j \sqcap C_k \equiv \bot$} peut impliquer l'existence :

- d'un problème de « Inconsistance logique ». En effet, le rajout de l'information $\{x \in Ci^I\}$ dans cette ontologie engendre une inconsistance (Ci^I étant une interprétation de Ci) ;
- des problèmes de « Ontologie inadaptée ». En effet, la possession de l'information : $IM \models Ci \sqsubseteq \exists R.\neg Cj$, $IM \models Ci \sqsubseteq \forall R.\neg C_k$ ou $IM \models not(C_j \sqcap C_k \equiv \bot)$ implique l'existence d'un cas de « Ontologie inadaptée » dans O ;
- des problèmes de « Contradiction sociale ». En effet, l'information : $SAP_O \models not\ (Ci \sqsubseteq \forall R.Cj)$, $SAP_O \models not\ (Ci \sqsubseteq \exists R.C_k)$ ou $SAP_O \models not\ (C_j \sqcap C_k \equiv \bot)$ implique l'existence d'un ou de plusieurs cas de « Contradiction sociale » dans O.

Pour formaliser les liens entre les problèmes de « Insatisfiabilité » et de « Contradiction sociale », nous avons considéré la formalisation des problèmes traités dans la littérature (section 3.3) et classés en « Insatisfiabilité » ou en « Contradiction sociale ». Parmi les 12 cas de « Insatisfiabilité » étudiés, seulement 8 cas sont différents (lignes grisées dans Tab. 3.3).

Problèmes d'insatisfaisabilité	Formalisation
P34-Anti-pattern AndisOR	$O \models C \sqsubseteq \exists R.(C_i \sqcap C_j)$, $C_i \sqcap C_j \equiv \bot$
P36 - Anti-pattern UniversalExistence	$O \models C \sqsubseteq \forall R.C_i$, $C \sqsubseteq \exists R.C_j$, $C_i \sqcap C_j \equiv \bot$
P37 - Anti-pattern EquivalenceIsDifference	$O \models C_i \equiv C_j$, $C_i \sqcap C_j \equiv \bot$
P45 - Partition error	$O \models C \sqsubseteq (C_i \sqcap C_j)$, $C_i \sqcap C_j \equiv \bot$
P47 - Having both a class and its complement as super condition	$O \models C \sqsubseteq (C_i \sqcap \neg C_i)$
P48 - Having a super condition that is assumed to be disjoint	$O \models C \sqsubseteq \neg \top$
P49 - Having a super condition that is an existential restriction that has a filler which is disjoint with the range of the restricted property	$O \models C \sqsubseteq \exists R.C_i$, $R.Range \equiv C_j$, $C_j \sqcap C_i \equiv \bot$
P50 - Having an universal restriction with Nothing as the filler and a must existing restriction along property relationships.	$O \models C \sqsubseteq \forall R.\bot$, $C \sqsubseteq \exists R.C_i$, $C_i \neq \bot$
P51 - Having more than allowed existential restrictions	$O \models C \sqsubseteq <2R.\top$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_j$, $C_i \sqcap C_j \equiv \bot$
P52 - Having a super condition containing conflicting cardinality restrictions	$O \models C \sqsubseteq > nR.\top$, $C \sqsubseteq <nR.\top$
P53 - Inconsistence from other ressources	$O \models C \sqsubseteq \exists R.C_i$, $C_i \sqsubseteq C_j$, $C_j \equiv \bot$
P54 - Having a super condition that is an existential restriction where the domain of the restricted property is disjoint with it	$O \models C_i \sqsubseteq \exists R.T$, $R.Domain \equiv C_j$, $C_i \sqcap C_j \equiv \bot$

Tab. 3.3 -Problèmes de « Insatisfiabilité » traités dans la littérature ([Harzallah, 2016])

Tab. 3.4 comprend la formalisation des 11 cas de problème de « Contradiction sociale ». Leur analyse a mis en évidence l'existence :

- dans 6 cas (*i.e.* dans P05, P19, P25, P27, P31et P41) d'un axiome d'égalité qu'un acteur social juge faux,
- dans 3 cas (*i.e.* dans P14, P15, P46) d'un axiome d'inclusion d'artefacts qu'un acteur social juge faux.
- dans 4 cas (*i.e.* P14, P15, P19, P41) sur 11, une correction du problème a été proposée dans la littérature (cf. axiomes soulignés dans Tab 3.4).

Cas de problème de contradiction sociale	Formalisation
P05 - Wrong inverse relationship	$O \models R^{-1}_i \equiv R_j$, $SAP_O \not\models R^{-1}_i \equiv R_j$
P14 - Misusing "Owl :allvaluesFrom"	$O \models C \sqsubseteq \forall R.C_i$, $SAP_O \not\models C \sqsubseteq \forall R.Ci$, $SAP_O \models \underline{C \sqsubseteq \exists R.C_j}$,
P15 - Misusing "not some" and "some not"	$O \models C \sqsubseteq \neg(\exists R.C_j)$, $SAP_O \not\models C \sqsubseteq \neg(\exists R.C_j)$, $SAP_O \models \underline{C \sqsubseteq \exists R.\neg C_j}$
P19 - Swapping intersection and union	$O \models R.range \equiv (C_i \sqcap C_k)$, $SAP_O \not\models R.range \equiv (C_i \sqcap C_k)$, $SAP_O \models \underline{R.range \equiv (C_i \sqcup C_k)}$
P25 - Define a relationship inverse to itself	$O \models R^{-1} \equiv R$, $SAP_O \not\models R^{-1} \equiv R$
P27 - Defining wrong equivalent relationship	$O \models R_i \equiv R_j$, $SAP_O \not\models R_i \equiv R_j$
P28 - Defining wrong symmetric relationship	$O \models Symmetric(R)$, $SAP_O \not\models Symmetric(R)$
P29 - Defining wrong transitive relationship	$O \models Transitive(R)$, $SAP_O \not\models Transitive(R)$
P31 - Defining wrong equivalent classes	$O \models C_i \equiv C_j$ $SAP_O \not\models C_i \equiv C_j$
P41 - DisjointnessOfComplement	$O \models C_j \equiv \neg C_j$, $SAP_O \not\models C_i \equiv \neg C_j$, $SAP_O \models \underline{C_i \sqcap C_j \equiv \bot}$
P46 - Semantic inconsistency	$O \models C_i \sqsubseteq R.C_i$, $SAP_O \not\models C_i \sqsubseteq R.C_i$

Tab. 3.4 - Problèmes de « Contradiction sociale » ([Harzallah, 2016])

Toutefois, en considérant les Tab. 3.3 et Tab 3.4, il n'y a pas de lien évident entre les formalisations des cas de ces deux types de problèmes. Pour cela, en se basant sur le principe qu'un problème d'insatisfiabilité « peut impliquer l'existence de » un problème de

« Contradiction sociale », nous n'avons considéré que la formalisation des cas d'« Insatisfiabilité » dans Tab 3.3. A partir de ce tableau, nous avons cherché à déterminer les axiomes qui pourraient correspondre les plus à un cas de contradiction sociale. Nous avons extrait, à partir de la formalisation des 8 cas d'insatisfiabilité, les axiomes pouvant représenter chacun une « Contradiction sociale ». Par exemple :

P36 où F(P36)={$C \sqsubseteq \forall R.Ci$, $C \sqsubseteq \exists R.C_j$, $C_i \sqcap C_j \equiv \bot$}, peut impliquer l'existence d'une ou de plusieurs « Contradictions sociales » parmi les suivantes :

- ($SAP_O \not\models C \sqsubseteq \forall R.Ci$),
- ($SAP_O \not\models C \sqsubseteq \exists R.C_j$),
- ou ($SAPo \not\models C_i \sqcap C_j \equiv \bot$).

Neuf axiomes types ont été extraits (on note cet ensemble A_{Ins}). Trois axiomes parmi eux représentent des problèmes connus de « Contradiction sociale » (*i.e.* P14, P15 et P31). Nous avons ordonné ces axiomes selon l'ordre décroissant, selon notre estimation, de la probabilité qu'un axiome corresponde à une contradiction sociale. Cette estimation dépend de la difficulté de la compréhension d'un axiome (critère évoqué dans [Roussey *et al.* 2010]) et de la pertinence de sa considération dans un processus de validation. Ce deuxième critère a été jugé en fonction de la rareté d'un axiome dans une ontologie et de son appartenance à plusieurs cas d'insatisfiabilité. L'ordre obtenu (noté Ord_{Ains}) est le suivant :

A1 : $C_i \sqcap C_j \equiv \bot$, A2 : $C_i \sqsubseteq \exists R.C_j$, A3 : $C_i \sqsubseteq \forall R.C_j$, A4 : $C_i \sqsubseteq <nR.C_j$, $n>1$, A5 : $C_i \sqsubseteq >nR.C_j$, $n>1$, A6 : $C_i \equiv C_j$, A7 : $R.Domain \equiv C_j$, A8 : $R.Range \equiv C_i$, A9 : $C_i \sqsubseteq C_j$.

Afin de minimiser l'intervention humaine pour l'identification des problèmes de contradiction sociale, nous avons proposé l'heuristique HPCS_APPI (Heuristique pour l'identification des Problèmes de « Contradiction Sociale » à l'aide des APPI). Cette heuristique se base sur l'idée que : *si on est proche d'une insatisfiabilité on est proche d'une « Contradiction sociale »* » [Harzallah, 2016]. Pour cela, nous avons extrait à partir des 8 cas différents d'insatisfiabilité, 13 anti-patrons partiels de l'insatisfiabilité (APPI) (cf. Tab.3.5). Un APPI est un patron de l'insatisfiabilité auquel on a enlevé un axiome (son axiome manquant).

Notre heuristique s'applique à une ontologie ne comprenant pas de problème d'insatisfiabilité (sinon il faut tout d'abord corriger les problèmes de ce type). Il cherche des contradictions sociales dans chaque occurrence d'un APPI (OAPPI), sans la transformer en une insatisfiabilité. Pour détecter une « Contradiction sociale » dans une OAPPI, on vérifie la validité des axiomes qui la composent. Une OAPPI n'est intéressante que s'il n'y a rien qui contredise son axiome manquant dans cette ontologie. A noter que l'APPI « $C_i \sqsubseteq C_j$ » n'est pas à prendre en compte car ses occurrences sont très fréquentes dans une ontologie.

Afin d'aider à cibler des cas de contradiction sociale, tout en minimisant l'intervention humaine, les APPI sont considérés selon l'ordre décroissant du nombre d'axiomes qui les composent et la validité de leurs axiomes est vérifiée selon l'ordre Ord_{Ains} défini ci-dessus.

Anti-patron partiel à 3 axiomes	Axiome manquant
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq <2R.\top$	$C \sqsubseteq \exists R.C_i$
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_i$	$C \sqsubseteq <2R.\top$
$C \sqsubseteq \exists R.C_i$, $C \sqsubseteq \exists R.C_i$, $C \sqsubseteq <2R.\top$	$C_i \sqcap C_j \equiv \bot$
Anti-patron partiel à 2 axiomes	Axiome manquant
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \forall R.Ci$,	$C \sqsubseteq \exists R.C_i$
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq \exists R.C_i$	$C \sqsubseteq \forall R.C_i$
$C \sqsubseteq \exists R.C_j$, $C \sqsubseteq \forall R.Ci$	$C_i \sqcap C_j \equiv \bot$
$Ci \sqsubseteq \exists R.T$, $R.Domain \equiv C_i$	$C_i \sqcap C_j \equiv \bot$
$C_i \sqcap C_j \equiv \bot$, $R.Domain \equiv C_i$	$Ci \sqsubseteq \exists R.T$
$C_i \sqcap C_j \equiv \bot$, $Ci \sqsubseteq \exists R.T$	$R.Domain \equiv C_j$
$C \sqsubseteq \exists R.Ci$, $R.range \equiv C_i$	$C_i \sqcap C_j \equiv \bot$,
$C_i \sqcap C_j \equiv \bot$, $R.range \equiv C_i$	$C \sqsubseteq \exists R.Ci$
$C_i \sqcap C_j \equiv \bot$, $C \sqsubseteq Ci$	$C \sqsubseteq Cj$
$C \sqsubseteq Ci$, $C \sqsubseteq Cj$	$C_i \sqcap C_j \equiv \bot$,
Anti-patron partiel à 1 axiome	Axiome manquant
$C_i \sqcap C_j \equiv \bot$	$C_i \equiv C_j$ ou $Ci \sqsubseteq C_j$
$Ci \equiv Cj$	$C_i \sqcap C_j \equiv \bot$
$C \sqsubseteq \exists R.(C_i \sqcap C_j)$	$C_i \sqcap C_j \equiv \bot$
$Ci \sqsubseteq Cj$	$C_i \sqcap C_j \equiv \bot$

Tab. 3.5 - Anti-patrons partiels de l'insatisfiabilité ([Harzallah, 2016])

Actuellement, nous cherchons à valider l'intérêt de l'ordre de vérification des axiomes de A_{ins} et celui de traitement des OAPPI pour la minimisation de l'intervention humaine. Nous avons cherché à les valider à l'aide d'un benchmark composé d'ontologies comprenant différents problèmes logiques et sociaux et la correction de ces ontologies. Cependant, d'après nos connaissances, ce type de benchmark n'existe pas encore.

Enfin, notre approche complète les travaux portant sur l'identification des cas susceptibles de correspondre à des problèmes (e.g. les travaux de Roussey *et al.* [2010] ou de Quazi&Qadir [2011]) et ceux qui font intervenir un acteur social pour l'aide à l'identification des problèmes.

3.3.1.4 Méthode pour la correction du problème de « Ontologie Plate »

Nous avons défini une méthode semi-automatique pour corriger le problème de « Ontologie plate » (*i.e.* une ontologie avec une profondeur faible parce qu'elle a un nombre important de concepts ayant la racine comme seul subsumant et ne subsumant aucun concept). C'est un des problèmes qui étaient flagrants dans la première version de l'ontologie ISTA3 construite avec Text2Onto. Ce problème peut être considéré en même temps comme une situation indésirable parce qu'il implique que l'ontologie obtenue soit difficilement réutilisable [Gangemi *et al.* 2006] et aussi comme une erreur parce qu'il nous prévient que cette ontologie pourrait être incomplète. En effet, le fait qu'une ontologie soit plate est dû potentiellement à l'absence de relations taxonomiques. Evidemment, certains raisonneurs peuvent proposer de nouvelles relations taxonomiques et améliorer la classification d'un concept. Cependant, ceci nécessite la présence dans l'ontologie d'axiomes autres que ceux qui définissent des relations taxonomiques.

Pour aider à corriger ce problème, nous avons proposé une méthode utilisant une ontologie noyau, pour identifier des nouvelles relations *est-un* entre les termes/concepts extraits et les concepts noyaux d'une ontologie.

Gruber [1993] a suggéré l'utilisation d'une ontologie noyau pour construire une ontologie du domaine. Certains travaux définissent ou réutilisent une ontologie noyau pour identifier et définir des concepts d'un domaine par spécialisation (*i.e.* en utilisant des relations taxonomiques). Par exemple, la majorité des OBO (Open Biomedical Ontologies) est définie à partir des ontologies

BFO (Basic Formal Ontology) et RO (Relation Ontology). D'autres travaux alignent l'ontologie noyau à une ontologie du domaine pour mieux définir les concepts de cette dernière et lui imposer une structure prédéfinie [Després&Szulman, 2007], [Burita *et al.* 2012]. Kutz&Hois [2012] pensent que l'utilisation d'une ontologie noyau améliore la modularité d'une ontologie. La modularité est liée au problème de « Ontologie Plate ». En effet, plus une ontologie est modulaire moins le nombre des concepts ayant la racine comme seul subsumant est important.

Peu de travaux ont envisagé d'utiliser une ontologie noyau pour corriger le problème d'« Ontologie plate » quand l'ontologie est construite d'une façon automatique à partir de textes.

Nous avons proposé une approche basée sur des règles inductives définies à partir d'une ontologie noyau afin d'aider à corriger le problème de « Ontologie plate ». Selon cette approche, à partir des relations entre les concepts noyaux et des relations taxonomiques validées dans une ontologie, des règles inductives sont appliquées pour classer des concepts extraits automatiquement, sous des concepts noyaux. Notre approche est basée sur le principe que si un concept a un contexte proche de celui d'un concept noyau alors ce dernier peut être un subsumant de ce concept. Elle complète les techniques présentées dans la section 2.4.1.2 (*e.g.* les techniques structurelles, les patrons lexicales).

Cette approche comprend trois étapes. Soient O une ontologie plate construire d'une façon automatique à partir de textes, C_i et C_j des concepts de O ; R_{kl} est une relation ente CC_k et CC_l ; R_{kl}, CC_k et CC_l sont des artefacts d'une ontologie noyau de O. La première étape est la génération d'une liste de synonymes des labels des concepts et relations noyaux. Cette liste peut être obtenue à partir de WordNet. Dans la suite, CCi fait référence au label du concept noyau CCi et à ses synonymes.

Dans la deuxième étape, les artefacts d'une ontologie noyau sont rajoutés à O d'une façon semi-automatique (les relations de cette dernière doivent être préalablement validées). En utilisant leurs labels, les concepts noyaux sont reliés automatiquement aux artefacts extraits de O. Par conséquent, une relation noyau entre deux concepts noyaux peut être héritée par deux concepts de O. Par exemple,. Si l'outil extrait les relations *est-un*(C_1,CC_1) et *est-un*(C_2,CC_2) alors la relation $R_{12}(C_1,C_2)$ est inférée.

La dernière étape est l'application de règles définies à partir de l'ontologie noyau. Ces règles sont appliquées lorsque leurs pré-conditions sont satisfaites. Un indice de confidence est associé à chacune d'entre elles. Il est estimé en fonction des informations requises par la règle elle-même (*i.e.* ses pré-conditions). Ces informations définissent la similarité des contextes des concepts impliqués dans une relation *est-un* inférée. Cet indice est un moyen empirique pour l'évaluation des relations inférées, il est à compléter par une évaluation humaine.

Règle1. Si $R_{kl}(CC_k,CC_l)$ est une relation de l'ontologie noyau, $R_{kl}(CC_k, C_j)$ (respectivement $R_{kl}(C_i,CC_l)$) est extraite ou inférée, il n'y a pas une relation *est-un* entre C_j (respectivement C_i) et un concept noyau, alors une relation *est-un* est suggérée entre C_j et CC_l (respectivement entre C_i et CC_k), avec un indice de confidence de 90%. Règle1 peut être définie comme suit :

$$R_{kl}(CC_k, CC_l) \wedge R_{kl}(CC_k, C_j) \Rightarrow est-un(C_i, C_j) \quad (3.1)$$

$$R_{kl}(CC_k, CC_l) \wedge R_{kl}(C_i, CC_l) \Rightarrow est-un(C_i, C_j) \quad (3.2)$$

Règle 2. Si $R_{kl}(CC_k,CC_l)$ est une relation de l'ontologie noyau, $R_{kl}(C_i, C_j)$ est extraite ou inférée, *est-un*(C_i, CC_k) (respectivement *est-un*(C_j, CC_l)) est extraite ou inférée et il n'y a pas de relation *est-un* entre C_j (respectivement C_i) et un concept noyau, dans ce cas, une relation taxonomique est suggérée entre C_j et CC_l (respectivement entre C_i et CC_k), avec un indice de confidence de 75%. La règle 2 peut être formalisée comme suit :

$$R_{kl}(CC_k, CC_l) \wedge R_{kl}(C_i, C_j) \wedge est-un(C_i, CC_k) \Rightarrow est-un(C_j, CC_l) \quad (3.3)$$

$$R_{kl}(CC_k, CC_l) \wedge R_{kl}(C_i, C_j) \wedge est-un(C_j, CC_l) \Rightarrow est-un(C_i, CC_l) \quad (3.4)$$

Règle3. Si $R_{kl}(CC_k,CC_l)$ est une relation de l'ontologie noyau, $R_{kl}(C_i, C_j)$ est extraite ou inférée et il n'y a pas de relation *est-un* entre C_i (respectivement C_j) et un concept noyau, les deux relation suivantes sont suggérées : *est-un*(C_i, CC_k) et *est-un*(Cj, CC_l) avec un indice de confidence de 50%. La règle 3 peut être formalisée comme suit :

$$R_{kl}(CC_k, CC_l) \wedge R_{kl}(C_i, C_j) \Rightarrow est-un(C_j, CC_l)$$

Nous avons expérimenté notre approche en utilisant Text2Onto, l'article de Navigli&Velardi [2004] et une ontologie noyau que nous avons construite. L'ontologie obtenue avant l'application de notre approche mais après sa correction (*i.e*. les termes inutiles et les relations taxonomiques et ad-hoc fausses ont été supprimés) comprend 118 concepts. Elle est plate (profondeur de 2) et elle est caractérisée par 64 concepts liés directement à la racine dont 34 ne subsument aucun concept. L'application de notre approche a permis de rajouter 28 nouvelles classifications faisant intervenir 20 concepts. La nouvelle ontologie a une profondeur de 3 et elle est caractérisée par 49 concepts liés directement à la racine dont 24 concepts ne subsument aucun concept. Le problème de « Ontologie Plate » a été réduit de 23% [Chulyadyo&Mittal, 2012].

La règle 3 de notre approche ressemble aux règles inductives définies dans plusieurs approches, par exemple l'approche définie dans [Faure&Nédellec, 1998a]. En effet, dans cette approche, à partir de données d'apprentissage, on détermine des patrons morphosyntaxiques pour certaines relations (dans notre approche, nous utilisons les patrons morphosyntaxiques des relations noyaux) et on applique des règles inductives définies à partir de ces patrons pour classer des termes/concepts. Par exemple, à partir du patron « <to extract> <sujet: tool> <from: resource> », on suppose que chaque fois qu'on rencontre le verbe « to extract » avec un sujet et un complément d'objet précédé par « from » alors le sujet est un « tool » et le complément d'objet est un « resource ». Poon& Domingos [2010] utilisent aussi des règles similaires aux nôtres définies à l'aide d'une analyse profonde de textes, et appliquent un raisonnement probabiliste pour inférer des nouvelles relations *est-un* (cf. section 2.1.4.2). Cependant, ces différentes approches n'utilisent pas les relations noyaux d'une ontologie, mais des relations à identifier.

3.3.2 Approche de construction et validation d'ontologie pour l'annotation

Dans cette section, nous présentons notre approche pour la construction et la validation d'ontologie lors de son utilisation pour l'annotation d'objets, un des domaines principaux d'application d'ontologie. L'annotation peut se faire avec une ontologie existante ou nécessite le développement d'une nouvelle ontologie adaptée au domaine des objets à annoter. Elle peut se faire à l'aide d'un concept, d'un ensemble de concepts ou d'instances, d'un graphe de concepts

ou d'instances, en utilisant ou non un patron prédéfini pour l'annotation. Certains travaux ont proposé de prendre en compte des patrons d'annotation dans l'étape d'enrichissement d'ontologie [Issac *et al.* 2005]. Ils ont discuté de la limite de la notation avec un formulaire prédéfini et ont proposé l'annotation avec un graphe de structure fixe dont les nœuds sont à définir avec les concepts d'une ontologie lors du processus d'annotation. Afin d'annoter un artefact logiciel, Amardeilh&Damljanovic [2009] proposent une phase de mapping manuel entre les termes extraits des documents portant sur cet artefact et une ontologie d'annotation et ensuite une phase de consolidation (ou de validation) de l'annotation et de peuplement d'ontologie. Cette phase de consolidation consiste à contrôler des contraintes telles que la non redondance, les restrictions de domaine et de co-domaine et les cardinalités.

Nédellec [2013] a considéré d'une façon conjointe l'annotation de textes et l'enrichissement ou le peuplement d'ontologie utilisée pour l'annotation : la mise à jour de cette ontologie est réalisée au cours du processus d'annotation. Elle a défini des consignes d'annotation (des guidelines) afin de limiter les risques d'incohérence. Une des perspectives de ces travaux est l'automatisation du maintien de la cohérence des annotations et de l'ontologie utilisées, des points de vue formel et social.

Dans nos travaux, nous nous sommes intéressés au processus de construction, enrichissement et validation d'une ontologie lors de son utilisation pour l'annotation d'objets. Cependant, il est à noter que nous ne considérons pas l'annotation comme un de nos thèmes de recherche. En effet, nous nous sommes intéressés au résultat de l'annotation d'un objet, réalisée à la main, pour définir sa sémantique, pour le comparer à d'autres objets (cf. chapitre suivant) ou pour enrichir une ontologie. En se basant, sur nos deux expériences dans ce domaine (UEMO pour l'annotation et la comparaison de constructs, et l'ontologie KIFANLO pour l'annotation et la comparaison des cartes cognitives), nous proposons une approche pour la construction et la validation de ce type d'ontologie qui généralise nos contributions dans le projet UEML [Harzallah *et al.* 2007], [Anaya *et al.* 2008], [Anaya *et al.* 2010], [Harzallah *et al.* 2012]. Cette approche permet de construire et d'enrichir une ontologie selon la nécessité de nouveaux artefacts pour l'annotation. Elle est dirigée par une ontologie noyau qui est elle-même définie en fonction des aspects à annoter.

La première étape de cette approche est la définition du méta-modèle d'annotation et d'une ontologie noyau. Un méta-modèle d'annotation va représenter les aspects clés des objets à annoter et des liens entre eux (*i.e.* une représentation structurelle ou fonctionnelle d'un objet). Ces aspects vont être utilisés pour comparer sémantiquement ces objets. Le méta-modèle est ensuite utilisé pour définir une ontologie noyau de l'ontologie d'annotation. Les entités du méta-modèle et les relations entre elles vont permettre de définir les concepts et les relations de cette ontologie noyau. Les contraintes de ce méta-modèle sont aussi à intégrer dans cette ontologie noyau afin d'améliorer sa formalisation. Fig. 4.2 du chapitre suivant illustre les liens entre le méta-modèle UEML pour l'annotation d'un construct et l'ontologie noyau d'UEMO. Dans ce méta-modèle, un construct pourrait être représenté par quatre types de phénomènes (ou aspects) : RepresentedThing, RepresentedProperty, RepresentedState et RepresentedTransformation reliés entre eux. Ce méta-modèle définit la représentation syntaxique d'un construct. Elle est ensuite annotée[32] par UEMO. Dans le projet KIFANLO, nous n'avons pas défini préalablement un méta-modèle d'annotation, les aspects à annoter et une ontologie noyau étaient définis au fur et à mesure de l'avancement et de la compréhension de l'objectif du projet (cf. chapitre précédent, section 2.5).

[32] Nous avons utilisé le mot « mapping » dans le lot de travail UEML

La deuxième étape de notre approche est l'enrichissement de l'ontologie noyau. Cette phase est nécessaire avant de commencer l'annotation. Cet enrichissement peut se faire par des experts (comme dans KIFANLO) ou en considérant des ontologies portant sur des domaines similaires (comme pour UEMO). La validation de cette première phase d'enrichissement est à réaliser à l'aide des experts du domaine et en respectant les axiomes de l'ontologie noyau.

La dernière étape est l'enrichissement et la validation de l'ontologie, au cours de son utilisation pour l'annotation de chaque objet. Afin de bien maîtriser ce processus et limiter le contenu de cette ontologie, nous ne l'enrichissons que par des artefacts nécessaires pour l'annotation en cours. Ainsi, son contenu est minimal mais évolutif : on peut l'enrichir autant de fois que des nouveaux objets à annoter le requièrent. Le processus d'enrichissement et validation est donc intégré au processus d'annotation. Il commence par (1) la définition structurelle et/ou fonctionnelle d'un objet avec le méta-modèle d'annotation, ensuite (2) cette définition est annotée avec l'ontologie d'annotation pour obtenir la définition sémantique de cet objet (son annotation). En cours de ce processus d'annotation, (3) on peut avoir besoin de rajouter des nouveaux concepts dans l'ontologie si on n'arrive pas à bien définir la sémantique de la représentation structurelle/fonctionnelle de cet objet avec cette ontologie.

Dans le projet UEML, nous avons défini des règles à appliquer pendant ou après ce processus d'annotation d'objet et d'enrichissement d'ontologie, afin d'éviter des problèmes dans la phase d'annotation ou dans celle d'enrichissement de l'ontologie. Ces règles sont généralisables à n'importe quel domaine d'application. Elles se basent sur la définition structurelle/fonctionnelle d'un objet et elles supposent que l'annotation doit la respecter. Par exemple, si la définition structurelle/fonctionnelle d'un objet comprend deux éléments reliés alors le résultat de l'annotation doit correspondre à deux concepts (ou deux ensembles de concepts) reliés. Si ce n'est pas possible, alors il y a un problème dans le résultat de l'annotation ou dans l'ontologie enrichie. Nous avons défini quatre types de règles : (1) des règles pour la validation de l'annotation, (2) des règles pour l'identification de l'absence éventuelle de relations dans une ontologie, (3) des règles pour l'identification d'axiomes de « disjointness » faux, et (4) des règles pour l'identification des relations ou axiomes (autres que des axiomes de « disjointness ») faux [Harzallah *et al*. 2012]. Nous présentons les trois derniers types de règles.

Règles pour l'identification d'un axiome de « disjointness » faux. La possibilité d'annoter un élément de la représentation structurelle/fonctionnelle d'un objet avec plusieurs concepts d'une ontologie implique que ces concepts ne doivent pas être disjoints, sinon cet élément n'a pas de sens. Trois règles (*i.e.* R8 (cf. Fig. 3.2), R9 et R10) définissent ce principe [Harzallah *et al.* 2012].

R8 Non-contradictory combination of ontology concepts.
If *Construct <described_by> RepresentedClass1&& RepresentedClass2&&…RepresentedClassn*
And RepresentedClass1 <represents> Concept1
And RepresentedClass2 <represents> Concept2
……
And RepresentedClassN <represents> ConceptN
And Concept1⊓Concept2⊓…ConceptN ≡⊥
Then *error (ontology or representation mapping is not well defined)*

Fig. 3.2 – R8 Non-contradictory combination of ontology concepts.

Règles pour l'identification de l'absence de relations. Le fait que l'annotation s'effectue suivant ou en respectant la définition structurelle/fonctionnelle d'un objet, requiert que le résultat

de l'annotation respecte les relations ou les contraintes de cette définition. Les règles de R11 à R15 définies dans [Harzallah *et al.* 2012] formalisent ce principe. Elles pourraient identifier, dans l'ontologie utilisée pour l'annotation, l'absence d'une relation de spécialisation entre deux concepts (R11) ou d'autres types de relations prédéfinies dans le méta-modèle d'annotation. La règle R13 (cf. Fig. 3.3) appliquée dans le projet UEML, suggère que si la représentation syntaxique d'un construct comprend le RepresentedPhenomena1 et le RepresentedPhenomena2 qui sont reliés par la relation « possesses » et annotés respectivement par les concepts C1 et P1, alors une relation «possesses » doit exister entre ces deux concepts, sinon, il y a un problème d'annotation de ce construct ou un problème dans cette ontologie.

R13 Non-contradictory representation mapping of related class-property
If *Construct <described_by> RepresentedClass1*
And *Construct <described_by> RepresentedProperty2*
And
RepresentedClass1 <possesses> RepresentedProperty2 (structurally related class-property)
And RepresentedClass1 <represents> Concept1
And RepresentedProperty2 <represents> Concept2
Then *Concept1 <possesses> Concept2*
Else *error (ontology or representation mapping is not well defined)*

Fig. 3.3 - R13 Non-contradictory representation

Règles pour l'identification des relations ou axiomes (autres que des axiomes de « disjointness ») faux. Ce type de règles permet d'identifier des relations ou axiomes dans l'ontologie d'annotation qui contredisent des relations dans la représentation structurelle/fonctionnelle d'un objet. Par exemple, si RepresentedPhenomena1 et RepresentedPhenomena2 d'un construct sont annotés chacun par C1 et C2 et il existe une relation « possesses » entre ces deux representatedphenomena alors que dans l'ontologie, on a un axiome qui interdit C1 et C2 d'être reliés par cette relation (*i.e.* C1⊓ possesses.C2 ≡⊥) alors il y a un problème dans l'annotation de ce construct ou l'axiome précédent est faux.

Les quatre types de règles ont été implémentés en SWRL pour améliorer l'étape d'annotation des constructs et d'enrichissement d'UEMO dans le travail de master de M. Abbas [2009].

3.4 Conclusion : Vers une approche semi-automatique de construction et validation intégrées basée sur une ontologie noyau

Dans ce chapitre, nous nous sommes intéressés à la validation d'ontologie. Nous avons abordé cette problématique de deux facettes. La première facette utilise les problèmes pouvant nuire à une ontologie, comme un moyen d'évaluer sa qualité et l'améliorer. Dans ce cadre, nous avons proposé une typologie originale des problèmes selon deux axes : les erreurs *versus* les situations indésirables et les problèmes sociaux *versus* les problèmes logiques. Nous avons également identifié des liens de causalité et de dépendance entre ces problèmes permettant d'optimiser leur identification.

La deuxième facette que nous avons abordée pour la validation d'ontologie est l'utilisation d'un score. Plus particulièrement, nous avons proposé une approche de validation d'ontologie intégrée au processus de son utilisation pour l'annotation.

Dans nos différents travaux sur les ontologies, nous nous sommes intéressés à l'utilisation d'une ontologie noyau pour gérer la complexité de la construction d'ontologie et de sa validation. Nous avons montré l'intérêt d'utiliser une ontologie noyau pour aider à corriger le problème de

« Ontologie Plate ». Nous avons montré également l'intérêt de son utilisation pour un enrichissement et une validation intégrées d'une ontologie lors de son utilisation pour l'annotation.

Enfin, nous avons utilisé une ontologie noyau dans nos différents projets de construction d'ontologie. Dans ces projets, la formalisation d'une ontologie noyau a paru un point clé pour obtenir une ontologie de bonne qualité et éviter un certain nombre d'erreurs.

En général, lors d'une construction manuelle d'une ontologie avec une approche ascendante ou descendante, on identifie des concepts du domaine de cette ontologie et on les positionne dans une hiérarchie. Dans ce cas, l'attention est portée à la sémantique de chaque concept et à sa position dans cette hiérarchie. Si on utilise une ontologie noyau, l'attention sera portée à la position d'un concept comme une spécialisation d'un concept noyau. Ce choix se fait généralement en fonction de la sémantique informelle/formelle des concepts noyaux et des concepts à positionner. Plus l'ontologie noyau utilisée est formelle plus son rôle pour l'amélioration de la qualité de l'ontologie obtenue sera important.

Au-delà d'une définition littéraire de chaque concept noyau, nous faisons deux recommandations pour la formalisation d'une ontologie noyau. Tout d'abord, nous recommandons sa formalisation, quand c'est possible avec les axiomes A1, A2, A3, A4, A5, A8 et A9 parmi ceux[33] présentés dans la section 3.3.1.3. A6 représente lui-même une situation indésirable et A7 ne se trouvera pas probablement dans une ontologie noyau, si on ne spécialise pas ses concepts. Ces axiomes font partie des patrons d'un nombre important de problèmes types dans une ontologie. L'existence de ce type d'axiomes dans une ontologie noyau permettrait d'éviter des problèmes fréquents ou de les détecter facilement. Plus précisément, ce type d'axiome permettrait quand on classifie un nouveau concept :

- d'engendrer une insatisfiabilité et donc d'aider à détecter une « Contradiction sociale ». Par exemple, si on rajoute un concept qui est une spécialisation de deux concepts noyaux disjoints (il s'agit d'un cas de A1), on crée une insatisfiabilité ce qui suggère que la classification de ce concept est fausse ;
- de s'approcher d'un patron partiel de « Contradiction sociale ». Par exemple, soit une ontologie noyau ayant deux concepts Cci et Ccj, une relation R et l'axiome Cci $\sqsubseteq \forall$R.Ccj (il s'agit d'un cas de A3), on rajoute C comme une spécialisation de Cci et tel que C $\sqsubseteq \exists R.C_k$, Ck étant un concept de l'ontologie à développer. Dans ce cas, C appartient à un patron partiel d'insatisfiabilité : C $\sqsubseteq \exists R.C_k$ et C $\sqsubseteq \forall$R.Ccj et le rajout de son axiome manquant engendrera une insatisfiabilité et suggérera, entre autre, de revoir la classification de C.

Notre deuxième recommandation pour la formalisation d'une ontologie noyau s'applique dans le cas où l'ontologie à développer est utilisée avec un méta-modèle pour l'annotation. Les axiomes et les contraintes du méta-modèle devraient être rajoutés dans une ontologie noyau pour le développement de l'ontologie pour l'annotation. Ensuite, nous pouvons appliquer les règles de la section 3.3.2 à cette dernière et valider son enrichissement.

Nos contributions citées dans ce chapitre peuvent s'articuler et former une approche semi-automatique de construction et validation intégrées, ayant les principes suivants :

1. Le développement d'une ontologie doit se baser sur une ontologie noyau formelle. Il doit être modulaire. Chaque concept noyau représentera la racine d'un module.

[33] Pour rappel, les 9 axiomes sont :A1= {$C_i \cap C_j = \perp$}, A2={ Ci $\sqsubseteq \exists R.C_j$ }, A3= { Ci $\sqsubseteq \forall$R.Cj}, A4={Ci $\sqsubseteq$ <nR.C_j and n>1}, A5={Ci $\sqsubseteq$ >nR.C_j and n>1}, A8={R.range=Ci}, A9={R.Domain=Cj}A6={Ci=Cj}, A7 ={Ci $\sqsubseteq$Cj}

2. La construction d’une ontologie noyau doit prendre en compte l’objectif de l’ontologie à développer. Cette ontologie noyau doit être formalisée, quand c’est possible, avec les axiomes cités dans la section 3.3.1.3. Si cette ontologie sert à développer une ontologie pour l’annotation, elle doit être alignée au méta-modèle d’annotation et ensuite enrichie et validée en appliquant les règles définies dans la section 3.3.2. Ces règles doivent aussi être appliquées à l’ontologie à développer après chaque enrichissement.
3. Enfin, nous proposons de suivre l’ordre d’identification des problèmes proposé dans 3.3.1.2 après chaque tâche du processus de conceptualisation ou après chaque étape d’enrichissement, pour valider l’ontologie. Nous recommandons de suivre l’ordre de validation des axiomes proposé dans la section 3.3.1.3. pour l’identification des problèmes de « Contradiction sociale » ou pour corriger un problème de « Insatisfiabilité ». La méthode de correction du problème de « Ontologie plate » peut être appliquée après chaque phase d’enrichissement et de validation.

La majorité des éléments de cette approche a été expérimentée chacun séparément. Nous souhaitons les expérimenter et les valider ensemble pour la conception d’une ontologie d’un cas réel.

Chapitre 4 : Cadre unifiant des mesures sémantiques de comparaison d'objets

Sommaire

4.1 Introduction

Avec l'évolution rapide et perpétuelle de l'environnement des organisations et le besoin de s'y adapter rapidement, la nécessité de s'interopérer vite, bien et au moindre coût est devenue primordiale pour ces organisations. Cependant, l'hétérogénéité des architectures de leurs applications ou des modèles et vocabulaires représentant leurs connaissances et données a rendu l'objectif d'interopérabilité difficile à atteindre. De même, l'hétérogénéité des données ou vocabulaires représentant les connaissances dans des bases de données, des documents ou sur le web a rendu la recherche d'information, la comparaison d'objets (*e.g.* gènes, documents, cartes géographiques)… difficiles à réaliser [Zargayouna *et al.* 2016], [Fritzsche&Gruninger, 2016]. En effet, les connaissances définies dans des vocabulaires différents peuvent être comprises d'une façon erronée. Afin de partager des connaissances d'une façon correcte entre plusieurs intéressés (e.g. agents, personnes, machines) et faire interopérer des systèmes utilisant des données et connaissances hétérogènes sans effort, trois approches ont été proposées : (1) une approche intégrée [Lenzerini, 2002] utilisant un schéma global et unique qui constitue une représentation virtuelle des données stockées dans différentes sources de connaissances, chaque source gardant quand-même son schéma de données ; (2) une approche unifiée en utilisant un méta-modèle ou un format pivot pour la mise en correspondance des données et connaissances et (3) une approche fédérée où la sémantique des données et connaissances est explicitée à l'aide d'ontologies. La troisième approche est intéressante parce que chaque organisation garde son vocabulaire pour exprimer ses connaissances ; parce que les correspondances sémantiques peuvent se faire entre les connaissances de plusieurs organisations et d'une façon bidirectionnelle et, enfin parce que cette approche permet de déterminer des correspondances entre les connaissances d'une nouvelle organisation et celles des organisations existantes, sans avoir besoin de tout redéfinir.

La mise en œuvre de l'approche fédérée pour l'interopérabilité des données et connaissances hétérogènes requiert :

- une ou plusieurs ontologies du domaine pour expliciter la sémantique de ces données et connaissances ;
- des mesures sémantiques de comparaison appliquées à ces ontologie(s) pour évaluer la proximité sémantique de ces données et connaissances et identifier des correspondances entre elles [Harzallah&Berio, 2015]. Dans le cas d'utilisation de plusieurs ontologies, des mesures d'alignement d'ontologies sont nécessaires [Shvaiko&Euzenat, 2005 ; 2013] ;
- le cas échéant, une architecture d'exploitation de ces ontologies dans une application pour l'interopérabilité des données et des connaissances [Wache *et al.* 2001], [Gayo, 2006], [Asgari *et al.* 2015], [Yang *et al.* 2016].

Dans ce chapitre, nous nous intéressons aux mesures sémantiques de comparaison de deux objets définis dans des vocabulaires différents. Un objet peut être un terme, un texte, un gène, une image, *etc.* De nombreuses mesures sémantiques ont été développées pour comparer la sémantique de deux objets pour des objectifs et des domaines variés. Nous pouvons notamment faire référence à la désambiguïsation de mots [Resnik, 1999], [Patwardhan *et al.* 2003], [Navigli&Velardi, 2005], [Navigli *et al.* 2011] ; à la détection et la correction de fautes d'orthographe [Budanitsky&Hirst, 2001] ; à la recherche d'images [Smeulders *et al.* 2000] ; à la recherche d'information [Hliaoutakis *et al.* 2006], [Sy *et al.* 2013], [Zargyouna *et al.* 2016] ou encore à diverses applications en biologie [Lord *et al.* 2003], [Pesquita *et al.* 2009], [Mazandu&Mulder, 2013], [Gaston *et al.* 2014]. [Mazandu *et al.* 2016].

Une mesure de comparaison sémantique évalue quantitativement la proximité ou la connectivité de deux objets, en explicitant la sémantique de chaque objet à l'aide d'une ou de

plusieurs ontologies. Comme présenté dans les chapitres précédents, le processus qui définit la sémantique d'un objet en utilisant une ontologie est généralement connu sous le nom de processus d'annotation sémantique [Prié&Garlatti, 2004], [Zavitsanos *et al.* 2010]. La façon dont un objet est annoté dépend de :

- la complexité des objets à comparer (*e.g.* deux termes versus deux documents) ;
- les aspects à prendre en compte dans la comparaison (*e.g.* la forme, la couleur, les fonctionnalités des objets) ;
- l'ontologie à utiliser, qui peut être plus au moins appropriée à l'annotation ;
- l'utilisation d'une ou de plusieurs ontologies.

Souvent, l'annotation d'un objet est réalisée à l'aide d'un seul concept d'une ontologie, mais parfois on préfère utiliser plusieurs concepts ou instances ou d'un graphe, appelé dans la suite graphe sémantique, dont les nœuds et les arcs sont étiquetés par les artefacts d'une ontologie. Zargayouna et Salotti [2004] ont annoté un terme par plusieurs concepts d'une ontologie, pour l'objectif d'indexation de documents. Desmontils et Jacquin [2001] ont classifié des documents sur le web en les annotant avec plusieurs concepts d'une ontologie : un document est annoté avec une liste de concepts associée à ses tags. Lorda *et al*. [2003] ont comparé des gènes (protéines) en les annotant avec exactement trois concepts, chacun appartenant à une des trois taxonomies[34] de l'ontologie des gènes (OG). Lee *et al*. [2008] et Mathur&Dinakarpandian [2012] ont comparé deux patients selon leurs maladies, en annotant chaque maladie par un concept d'une ontologie. Pesquita *et al.* [2008] et Mazandu *et al.* [2016] comparent deux protéines en annotant chacun par un ensemble de concepts. Baziz *et al.* [2004] ont annoté des documents, chacun par un vecteur de concepts pour les comparer. Pour la recherche d'information dans des documents audiovisuels, Issac *et al.* [2005] ont annoté un document à l'aide d'un patron pré-défini sous la forme d'un graphe sémantique. Pour l'objectif de l'interopérabilité des modèles d'entreprise, Anaya *et al.* [2010] ont annoté les constructs des langages de modélisation d'entreprise, chacun par un graphe sémantique.

La majorité des mesures sémantiques proposées compare des objets annotés chacun par un concept d'une même ontologie. Certaines d'entre elles ont été étendues à la comparaison d'objets annotés chacun par un ensemble de concepts. Souvent, ces mesures comme celles définies en analyse des données, sont basées sur le principe de Shannon en théorie de l'information : la comparaison de deux individus ou objets dépend de l'information commune quantifiée apportée par ces deux individus ou objets et celle distinctive apportée par chacun d'entre eux. Cependant, il est difficile de comprendre les liens entre ces mesures et d'en choisir une pour une application et un contexte donnés. En plus, une mesure peut changer de résultats si l'ontologie utilisée subit une suppression ou un rajout de concepts. Une mesure peut aussi changer de précision si on change le jeu de données sur lequel elle est appliquée.

Par ailleurs, nous avons constaté l'absence de mesures sémantiques basées sur le principe de Shannon et comparant des objets annotés chacun par un graphe sémantique. Evidemment, il y a plusieurs travaux portant sur la comparaison des graphes de grande ou de petite taille avec ou sans étiquette.

Dans nos travaux, nous avons proposé un cadre pour définir d'une façon unifiée et paramétrable trois familles de mesures sémantiques, chacune utilisant un type d'annotation : par un concept, par un ensemble de concepts ou par un graphe sémantique. En effet, tout en respectant le principe de Shannon, les trois familles de mesures ont la même forme et quantifient

[34] i.e. une taxonomie des fonctions des molécules, une taxonomie des processus biologiques et une taxonomie des composants cellulaires.

l'information apportée par un objet par une approximation du contenu informationnel de son annotation. Des approximations du contenu informationnel d'un concept ont été proposées dans la littérature, mais sans expliquer leurs liens avec la définition originelle du contenu informationnel. Nous avons proposé des approximations du contenu informationnel d'un concept, d'un ensemble de concepts et d'un graphe sémantique, tout en prenant en compte cette définition.

L'aspect paramétrable de notre cadre permet de faciliter le choix d'une mesure qui s'adapte au mieux à l'objectif et aux besoins d'une application, de bien comprendre la sémantique de ses résultats et de pouvoir les comparer aux résultats d'autres mesures.

Dans ce chapitre, nous présentons tout d'abord l'exploitation des mesures sémantiques dans l'approche UEML (Unified Enterprise Modelling Language) pour l'interopérabilité des langages de modélisation (section 4.2). Cette approche se base sur la détermination des correspondances sémantiques entre constructs des langages différents, constructs annotés avec UEMO. Nous utilisons un cas d'exemple de cette approche pour illustrer nos contributions en mesures sémantiques de comparaison.

Nous présentons ensuite les trois familles de mesures sémantiques de comparaison que nous avons définies, tout en les positionnant par rapport aux travaux existants (sections 4.3, 4.5 et 4.6). Nous détaillons en particulier la troisième famille de mesures qui permet de comparer deux objets annotés chacun par un graphe sémantique. Dans la section 4.4, nous présentons notre approche pour l'approximation du contenu informationnel, élément principal et commun à la définition de ces trois familles de mesures.

Nous présentons dans la section 4.7 la variabilité des résultats des mesures sémantiques ayant chacune une approximation différente du contenu informationnel. Nous traitons également la variabilité des résultats d'une mesure donnée avec l'évolution de l'ontologie à laquelle elle est appliquée. Nous présentons dans la section 4.8 notre cadre pour la définition des mesures sémantiques et ses paramètres avant de clore ce chapitre par des perspectives d'amélioration de nos contributions.

4.2 Comparaison sémantique d'objets dans l'approche UEML

L'interopérabilité est la capacité d'au moins deux systèmes à échanger des informations et à accéder réciproquement à leurs fonctionnalités. Elle est devenue un enjeu majeur si on veut faire communiquer vite et à la volée des applications hétérogènes et changeantes dans un environnement dynamique, de façon économique et flexible. Dans le Rex INTEROP, on a proposé, une approche dite « Interopérabilité Guidée par les Modèles » (Model Driven Interoperability (MDI)) qui traite la question des échanges entre applications depuis leur niveau Business. Cette approche est dérivée de l'approche MDA (Model Driven Architecture) définie et adoptée par l'OMG (Object Management Group). Elle vise à promouvoir l'utilisation de modèles et de leurs transformations pour concevoir et implanter différents systèmes et les faire interopérer.

Dans le lot de travail UEML du Rex INTEROP, nous nous sommes intéressés à la problématique d'interopérabilité des systèmes produisant ou traitant des modèles d'entreprise, des systèmes de modélisation ou de simulation des processus d'entreprise, par exemple. Nous avons proposé une architecture s'intégrant dans l'approche MDI et utilisant l'approche UEML pour l'interopérabilité des langages. Dans cette architecture trois types d'ontologie ont été identifiés. Le premier type d'ontologie est une ontologie de langage pour aider à comparer les constructs des langages (*i.e.* éléments de base d'un langage, par exemple : Activité, Classe, Objet du langage UML) avec lesquels les modèles des deux systèmes sont définis et à identifier des

correspondances sémantiques entre eux. Le deuxième type d'ontologie est une ontologie de domaine pour identifier des relations sémantiques entre les modèles manipulés par les deux systèmes, en utilisant comme contraintes les correspondances entre les langages. Le dernier type d'ontologie est une ontologie de services pour identifier des correspondances sémantiques entre des services offerts par un système et des services demandés par l'autre système, tout en utilisant les correspondances identifiés entre les modèles (cf. Fig. 4.1).

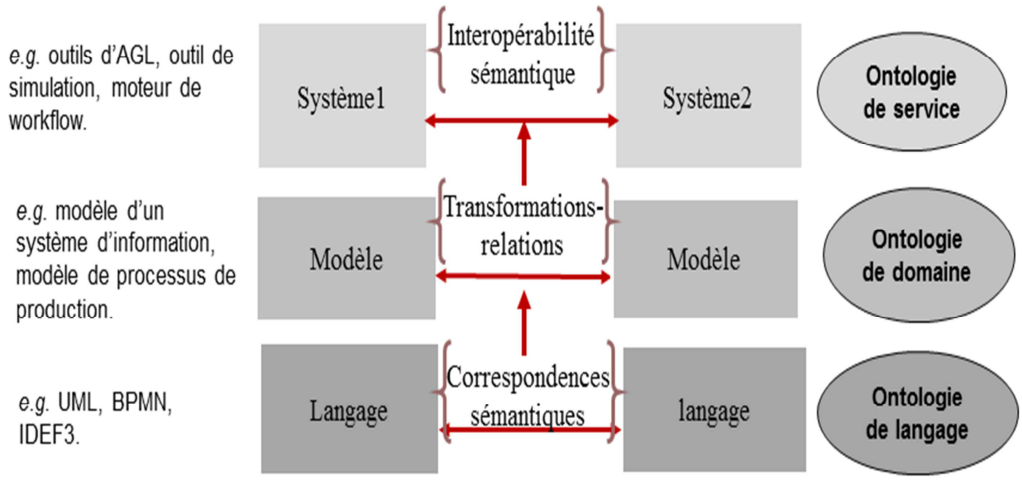

Fig. 4.1 - Architecture pour l'exploitation des ontologies pour l'interopérabilité des systèmes

Plusieurs approches se sont intéressées à interopérer des modèles ou des langages. Certaines se sont focalisées sur la transformation de modèles [Bézivin, 2006], [Benelallam *et al.* 2015]. Elles sont bilatérales et unidirectionnelles et ne considèrent que des aspects syntaxiques de l'interopérabilité alors que les problèmes d'interopérabilité sont dus aussi à des aspects sémantiques. D'autres approches ont traité les relations entre langages deux à deux, en analysant manuellement (par des experts) leurs correspondances. Leurs résultats sont peu réutilisables, vu qu'ils ne peuvent pas être étendus à un groupe de langages. La plupart des analyses ontologiques des langages a porté sur un seul langage et parfois sur un seul couple de constructs *e.g.* [Rohde, 1995], [Green&Rosemann, 2000], [Opdahl&Henderson-Sellers, 2002], [Evermann *et al.* 2005], [Mazak &Huemer, 2015]. Cependant, ces contributions n'ont pas formalisé leur processus d'utilisation d'une ontologie. Leurs résultats sont souvent différents, et parfois contradictoires pour le même langage [Gehlert&Esswein, 2007].

Nous nous sommes intéressés à la première couche d'interopérabilité, en l'occurrence l'interopérabilité des langages de modélisation. Nous avons contribué au développement de l'approche UEML pour la modélisation intégrée des langages de modélisation d'entreprise. L'approche UEML se base sur une représentation ontologique des langages. Elle propose de définir une représentation structurelle des langages à l'aide du méta-modèle UEML (cf. Fig.4.2) et ensuite d'annoter cette représentation avec UEMO, suivant un processus d'incorporation de langage [Harzallah *et al.* 2012]. Des mesures sémantiques sont ensuite appliquées à ces annotations (appelées également représentations ontologiques) pour les comparer et déterminer des correspondances sémantiques entre les constructs des langages.

Comme présenté dans le chapitre précédent, pour définir une représentation structurelle d'un langage, la structure syntaxique de chaque construct est définie en fonction de quatre types de phénomènes : RepresentedThing, RepresentedProperty, RepresentedSate et RepresentedTransformation (1ère couche dans Fig.4.2). Cette représentation structurelle joue le rôle d'un méta-modèle pour l'annotation d'un construct. Pour chaque type de phénomène, une taxonomie est définie dans UEMO. En effet, UEMO est composée de quatre taxonomies reliées par des relations non taxonomiques : Class, Property, State and Transformation (2ème couche dans Fig. 4.2). La représentation structurelle de chaque construct est annotée avec UEMO en deux étapes : tout d'abord, un construct est annoté avec un unique concept de cette ontologie qui est le phénomène central que représente ce construct. Ensuite, les autres éléments de sa représentation structurelle sont annotés avec des concepts de cette ontologie pour définir la sémantique de chacun d'entre eux. Seules les taxonomies Class et Property sont considérées car elles sont les seules taxonomies assez développées pour que leur exploitation soit pertinente.

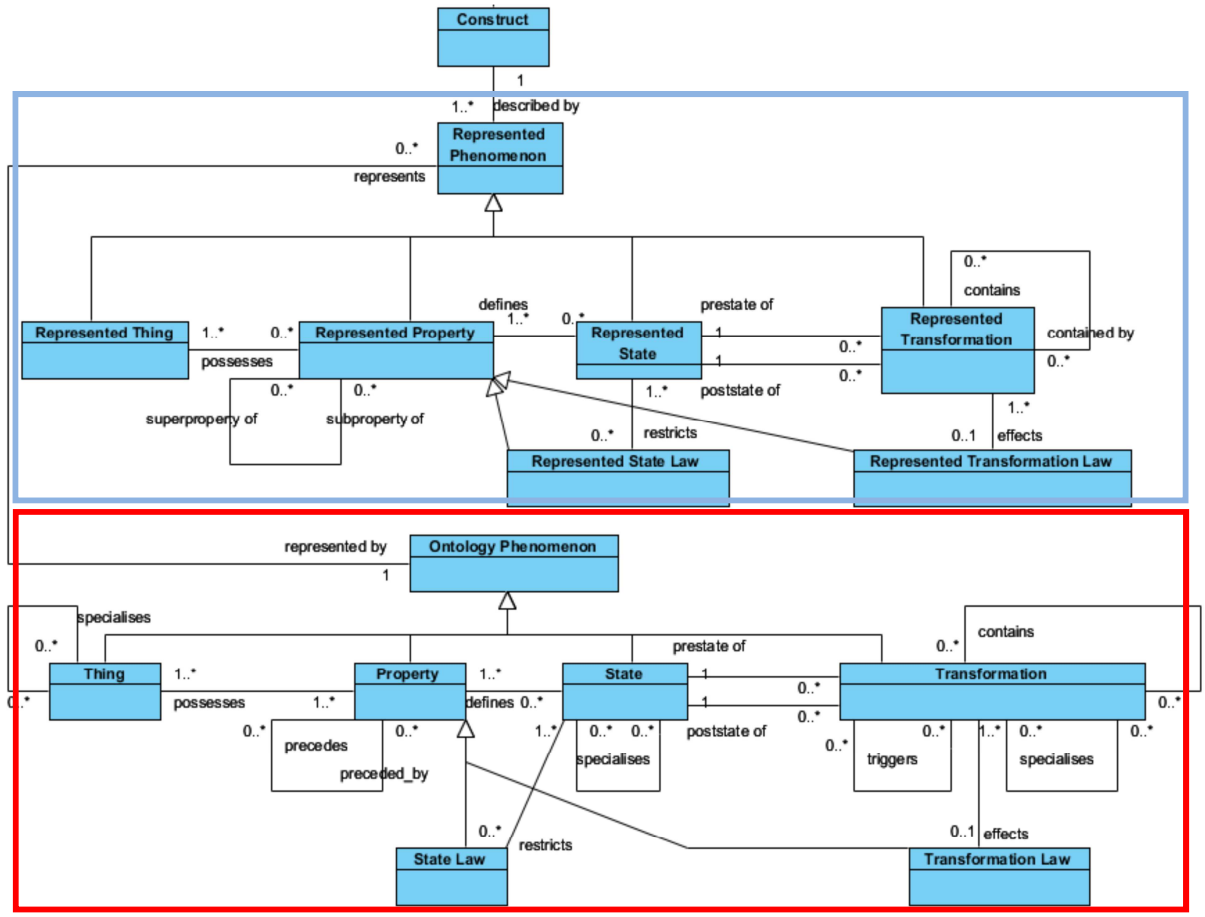

Fig. 4.2 - Le méta-modèle de UEML (cadre du haut) et l'ontologie noyau d'UEMO (cadre du bas) [Anaya *et al.* 2010]

Fig. 4.3 illustre les étapes d'incorporation dans l'approche UEML de trois exemples de constructs (CT1, CT2, CT3) que nous utilisons pour comparer les familles de mesures sémantiques que nous avons proposées.

Dans Fig. 4.3, la structure syntaxique de chacun de ces constructs est définie dans la partie « construct representation » ; les concepts subsumés par C0 (y compris C0 lui-même) représentent des concepts de la taxonomie Class ; les concepts subsumés par P0 (y compris P0 lui-même) représentent des concepts de la taxonomie *Property ; et* les lignes en pointillés indiquent l'alignement des composants de la structure syntaxique d'un construct aux artefacts de l'ontologie.

Le graphe sémantique d'un construct (ou sa représentation ontologique) est défini dans la dernière partie de Fig. 4.3 (Mapping graph result). Il s'agit du graphe associé à un construct, obtenu en représentant la sémantique d'un construct en termes de concepts et de relations d'UEMO, concepts et relations pouvant être répétés plusieurs fois.

Suite à l'incorporation d'un construct dans UEML, on peut distinguer trois types d'annotations pouvant être utilisés pour la comparaison de constructs :

- Un construct peut être comparé en utilisant seulement son annotation avec le concept qui est aligné à son phénomène central [Harzallah *et al*. 2012]. Dans ce cas, l'annotation A_c des constructs de Fig. 4.3 par un seul concept est :

  A_c (CT1) = C9, A_c(CT2) = C9, et A_c(CT3) = C9

- Un construct peut être comparé en utilisant son annotation avec les concepts qui sont alignés à sa structure syntaxique. Dans ce cas, l'annotation A_s des constructs de Fig. 4.3 par un ensemble de concepts est :

  A_s(CT1)= {C9, P4, P6}, A_s(CT2)= {C9, P6} et A_s(CT3)= {C9, P6, P7}

- Enfin, un construct peut être comparé en utilisant la totalité de l'alignement de sa structure syntaxique. Dans ce cas, l'annotation A_g des constructs de Fig. 4.3 par un graphe sémantique est :

  A_g(CT1)= {n_{11}(C9), n_{12}(P6), n_{13}(P4), e_{11}(n_{11}, n_{12}, possesses), e_{12}(n_{11}, n_{13}, possesses)}
  A_g(CT2)= {n_{21}(C9), n_{22}(P6), e21(n_{21}, n_{22}, possesses)}
  A_g(CT3) = {n_{31}(C9), n_{32}(P6), n_{33}(P7), e_{31}(n_{31}, n_{32}, possesses), e_{32}(n_{31}, n_{33}, possesses)}

  où nij (Ck) et e_{ij} (nij, nik, Ri) sont respectivement les nœuds et les arcs du graphe considéré, Ck et Ri sont respectivement des concepts et des relations de l'UEMO et représentent les labels respectivement des nœuds et des arcs de ce graphe.

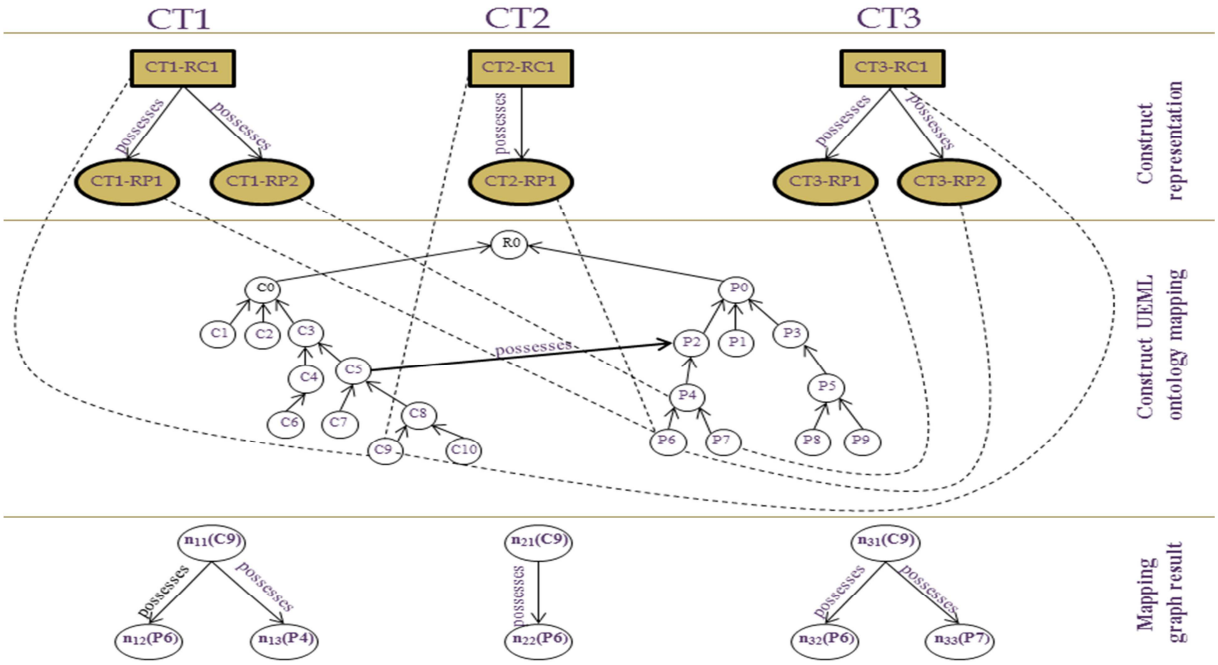

Fig. 4.3 - Incorporation des constructs des langages dans l'approche UEML [Harzallah *et al.* 2015]

Fig. 4.4 montre le graphe sémantique du construct « Activity » du langage UML, résultat de son incorporation dans l'approche UEML

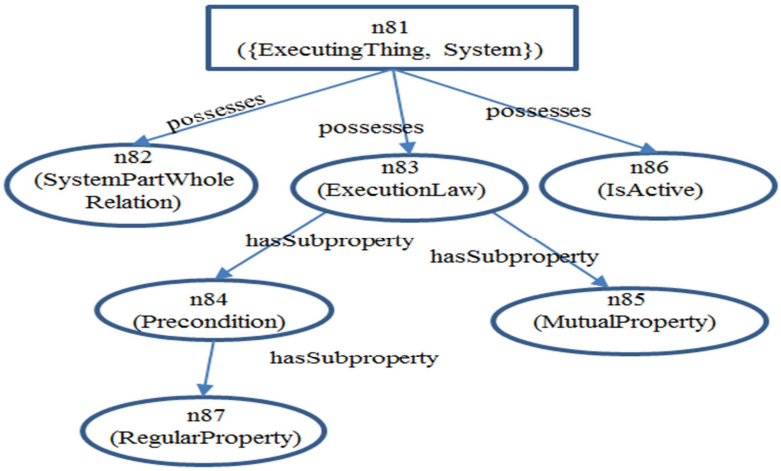

Fig. 4.4 - Graphe sémantique du construct « Activity » de UML

4.3 Mesures sémantiques de comparaison de deux concepts

Bien évidemment, comparer deux objets annotés chacun par un concept d'une ontologie revient à comparer les deux concepts qui les annotent. De nombreuses mesures sémantiques pour la comparaison de deux concepts ont été définies dans la littérature. La majorité de ces mesures

est basée sur le principe de Shannon en théorie de l'information qui considère qu'une mesure de comparaison de deux objets dépend de l'information commune apportée par ces deux objets et de l'information distinctive que chacun d'eux apporte par rapport à l'autre. Ce principe a été utilisé en psychologie dans le modèle de contraste de Tversky [1977]. Considérons deux ensembles A et B et une fonction f qui permet d'associer une valeur numérique à un ensemble (par exemple sa cardinalité). Le modèle de contraste de Tversky a été défini par M1(A,B) comme suit :

$$M1(A, B)= \theta\, f(A \cap B) - \alpha f(A-B) - \beta f(B-A) \qquad (4.1)$$

où f est une fonction qui quantifie un ensemble et β, α et θ sont trois réels positifs. Le modèle de contraste de Tversky normalisé a été défini par M2(A,B) comme suit :

$$M2(A,B) = \frac{f(A \cap B)}{\alpha f(A - B) + \beta f(B - A) + f(A \cap B)} \qquad (4.2)$$

Par ailleurs, Rodriguez&Egenhofer [2003], Blanchard *et al.* [2007] et Mazandu *et al.* [2016] ont classé la majorité des mesures sémantiques en trois familles, selon le type de données utilisé en entrée :

- Mesures utilisant seulement la structure taxonomique d'une ontologie, par exemple la profondeur d'un concept dans une taxonomie, le nombre de ses fils, le chemin le plus court entre deux concepts via leur subsumant commun le plus spécifique [Rada *et al.* 1989], [Sussna, 1993], [Wu&Plamer, 1994], [Blanchard *et al.* 2006b] ;
- Mesures utilisant la structure taxonomique d'une ontologie et une intension *i.e.* les caractéristiques intrinsèques des artefacts de cette ontologie (*e.g.* les attributs des concepts, les relations, leur cardinalité, leur domaine et leur co-domaine) [Ganesan *et al.* 2003], [Shvaiko& Euzénat, 2013] ;
- Mesures utilisant la structure taxonomique d'une ontologie et une extension (*e.g.* des instances des concepts de cette ontologie [Resnik, 1995], les occurrences des concepts dans un corpus du domaine de cette ontologie [Resnik, 1995], [Jiang&Conrath, 1997], [Lin, 1998], des données ou textes annotés par les concepts de cette ontologie [Mazandu *et al.* 2016]). Ces mesures utilisent une approximation du contenu informationnel d'un concept (noté IC(C)). En effet, pour quantifier l'information qu'apporte un concept C d'une ontologie, Resnik [1995] a proposé d'évaluer son contenu informationnel par l'entropie de Shannon tel que IC(C) = −log P(C), où P(C) est la probabilité qu'une instance donnée appartienne à l'ensemble des instances de C. Cependant, l'ensemble des instances de chaque concept d'une ontologie ne peut pas être complétement connu. Plusieurs approximations du contenu informationnel ont été proposées (cf. section 4.4).

En se basant sur une étude approfondie de l'état de l'art, nous avons proposé une forme unifiée pour les mesures sémantiques de comparaison de deux concepts, à partir du modèle de contraste de Tversky normalisé (4.2) [Blanchard *et al.* 2008b]. Elle est définie comme suit :

$$M(C1, C2) = \frac{ICa^{\cap}(C1, C2)}{\alpha ICa^{-}(C1, C2) + \beta ICa^{-}(C2, C1) + ICa^{\cap}(C1, C2)} \qquad (4.3)$$

C1 et C2 sont deux concepts d'une taxonomie. ICa est une approximation du contenu informationnel d'un concept. $ICa^{\cap}$ et ICa^{-} sont des approximations respectivement de la

quantification de l'information commune apportée par deux concepts et de l'information distinctive apportée par un concept par rapport à l'autre. Elles sont définies comme suit :

$$ICa^{\cap}(C1,C2) = ICa(C_{mscs}) \quad (4.4)$$

$$ICa^{-}(C1,C2) = ICa(C1) - ICa^{\cap}(C1,C2) \quad (4.5)$$

C_{mscs} est le subsumant commun le plus spécifique de C1 et C2.

L'état de l'art de Choi *et al.* [2010] a confirmé que cette forme unifiée reste assez générale pour couvrir ou être proche de celles des mesures existantes. Elle est suffisamment générique pour s'adapter à différentes problématiques de comparaison. α et β sont deux réels positifs qui pondèrent l'importance de l'information distinctive apportée par chaque concept et permettent ainsi de définir un type de mesure pour un objectif de comparaison requis (*e.g.* une similarité, une inclusion). Par exemple, quand α = β, les mesures obtenues sont des mesures symétriques permettant l'évaluation de la similarité de deux concepts. Quand α =0, l'ensemble de ces mesures sont des mesures asymétriques permettant l'évaluation de l'inclusion de C1 dans C2 ou de l'intersection de ces deux concepts.

Si on applique les formules 4.4 et 4.5 aux constructs CT1, CT2 et CT3 de la section 4.2, en utilisant l'annotation avec un seul concept, on obtient :

$ICa^{\cap}(CT1,CT2) = ICa^{\cap}(CT1, CT3) = ICa^{\cap}(CT2, CT3) = ICa(C9)$
$ICa^{-}(CT1, CT2) = ICa^{-}(CT1, CT3) = ICa^{-}(CT2, CT4) = ICa(C9) - ICa(C9) = 0$

4.4 Approximation du contenu informationnel d'un concept

« *La théorie de l'information vise à quantifier et qualifier le contenu en information présent dans un ensemble de données. Pour Shannon, l'information présente un caractère essentiellement aléatoire et donc incertain. Cette incertitude mesure l'information : plus une information est incertaine plus elle est intéressante* ». Resnik [1995] s'est basé sur cette théorie et a défini le contenu informationnel d'un concept par IC(C) = −log P(C), où P(C) est la probabilité qu'une instance donnée appartienne à l'ensemble d'instances de C.

Le contenu informationnel est une fonction monotone (*i.e.* si Ci *est-un* Cj alors IC(Ci)> IC(Cj)) qui détermine le niveau de spécificité d'un concept. L'information apportée par un concept est en partie apportée par ses ancêtres :

$$P(Ci) = P(Cj).\rho ij \quad (4.6)$$

où ρij est la probabilité qu'une instance quelconque appartienne à l'extension du concept *Ci* sachant qu'elle appartient à celle du concept Cj, Cj étant un ancêtre de Ci. Ceci implique que :

$$IC(Ci) = IC(Cj) - \log(\rho ij) \quad (4.7)$$

En effet, soient Ei et Ej respectivement les ensembles d'instances de Ci et Cj :

$$\rho ij = P(Ei/Ej) = P(Ei \cap Ej)/P(Ej) = P(Ei)/P(Ej) \quad (4.8)$$

Toutes les instances des concepts sont difficilement disponibles, il est intéressant de définir une approximation du contenu informationnel en utilisant les données disponibles sur les artefacts de l'ontologie.

Resnik [1995] a proposé une approximation du contenu informationnel, en utilisant un corpus représentatif de l'ontologie utilisée : les instances d'un concept sont remplacées par ses occurrences et celles de ses ancêtres. La qualité de cette approximation dépend de la qualité du corpus et de l'ontologie utilisés. La qualité d'un corpus dépend par exemple du nombre de concepts de l'ontologie mentionnés dans ce corpus. La qualité d'une ontologie peut dépendre du nombre de concepts ou de relations de subsumption manquants dans cette ontologie.

D'autres travaux en bioinformatique ont défini des approximations d'IC en remplaçant les instances d'un concept par les protéines qui annotent un concept. En effet, un concept de l'ontologie des gènes (OG) peut être utilisé pour annoter un ou plusieurs protéines dans une base de données. On considère que plus un concept est utilisé pour l'annotation de protéines moins il est spécialisé. Trois approximations ont été proposées pour le calcul de IC : (1) la fréquence relative d'un concept dans l'annotation de protéines [Mazandu *et al.* 2016], (2) la fréquence relative d'un concept et ses ancêtres dans l'annotation de protéines [Wang *et al.* 2007] ou (3) la fréquence relative d'un concept et ses descendants dans l'annotation de protéines [Zhang *et al.* 2006]. Il est clair que ces trois approximations ne donnent pas la même valeur pour le contenu informationnel d'un concept. La sémantique de chacune de ces trois approximations et son lien avec la définition originelle du contenu informationnel ne sont pas expliqués. Il est difficile de bien les exploiter et de comprendre les résultats des mesures sémantiques qui les utilisent. Par ailleurs, l'utilisation de la première approximation induit qu'un concept qui annote plus de protéines que son ancêtre possède un contenu informationnel inférieur à celui de ce dernier, ce qui n'est pas correct (cf. section 4.3))

Plusieurs approches ont proposé une alternative à l'utilisation d'un corpus de textes ou d'un autre type d'extension d'une ontologie [Seco *et al.* 2004], [Zhou *et al.* 2008], [Sanchez *et al.* 2011]. Cette alternative consiste à redéfinir le contenu informationnel en considérant uniquement la structure taxonomique d'une ontologie. Elle évite les inconvénients de l'utilisation d'un corpus lors du calcul du contenu informationnel. En effet, un corpus peut contenir des termes dont il faut clarifier le sens dans un domaine spécifique. En plus, l'utilisation de corpus différents pour un même domaine pourrait donner des valeurs différentes pour le contenu informationnel d'un même concept [Mazandu *et al.* 2013b]. Cette alternative se base sur l'intuition selon laquelle l'information principale extraite par la méthode du calcul du contenu informationnel est en grande partie inhérente à la structure taxonomique. Pour cette alternative, les auteurs ont défini une fonction f pour approximer le contenu informationnel tel que ICa=-log (f), où f prend en compte certaines caractéristiques structurelles d'un concept dans une taxonomie (*e.g.* sa profondeur, le nombre de ses fils, le nombre de ses ancêtres) qui influencent son contenu informationnel. Toutefois, ils n'ont expliqué ni le lien de chaque proposition avec la formule originelle du contenu informationnel, ni la sémantique de chacune et son intérêt par rapport à une autre.

D'autres travaux se sont intéressés aussi à définir le contenu informationnel sans corpus mais en considérant les autres relations dans une ontologie en plus de la relation de subsumption [Suzanna, 1997], [Buggenhout&Ceusters, 2005], [Pirro&Euzenat, 2010]. Ces travaux comme les précédents, n'ont pas fait le lien entre leur proposition et la définition originelle du contenu informationnel. Dans cette section, nous ne considérons pas ces travaux, étant donné que nous ne nous sommes focalisés que sur les mesures sémantiques appliquées à une taxonomie.

Comme les travaux précédents, nous avons trouvé intéressant de calculer le contenu informationnel sans avoir recours à une extension d'une ontologie. Nous avons proposé des approximations du contenu informationnel, en fonction uniquement de l'information structurelle associée à une taxonomie, mais en réutilisant la formule originelle du contenu informationnel. Il

s'agit d'une vraie approximation en faisant des hypothèses sur la distribution des instances sur les différents concepts de l'ontologie. Ceci permet de bien comprendre la sémantique de chaque approximation du contenu informationnel et ses résultats.

Dans la thèse de E. Blanchard [2008], nous avons défini trois approximations de la probabilité qu'une instance donnée appartienne à l'ensemble d'instances d'un concept C (i.e. P(C), IC=-log(P(C)) : P_p, P_s, et P_g.

Soient O une ontologie, Père(c_i) le subsumant direct de c_i et Fils(c_i) les subsumés directs de c_i.

Pour l'approximation P_p (formule 4.9), on suppose que le nombre d'instances d'un concept est divisé par k à chaque spécialisation, k étant un entier fixé supérieur ou égal à 2.

$$P_p(c_i) = \frac{P_p(\text{Père}(c_i))}{k^{P_i}} \qquad (4.9)$$

où p_i est la profondeur de ci et k un entier fixé supérieur à 1

Pour l'approximation P_s (formule 4.10), on suppose que la distribution du nombre d'instances est uniforme sur l'ensemble des fils de chaque concept.

$$P_s = \frac{P_s(\text{Père}(c_i)}{|\text{Fils}(\text{Père}(c_i)|} \qquad (4.10)$$

Pour l'approximation P_g (formule 4.11), on suppose que la distribution du nombre d'instances est uniforme sur l'ensemble des feuilles de la taxonomie.

$$P_g(c_i) = \frac{P_g(\text{Père}(c_0))}{|C_l|} \quad \text{si } c_i \in Cl, Cl \text{ étant l'ensemble des feuilles de la taxonomie}$$

$$\text{sinon } P_g(c_i) = \sum_{c_x \in \text{Fils}(c_i)} P_g(c_x) \qquad (4.11)$$

P_s et P_g sont complémentaires, P_s dépend de la structure de O entre la racine c_0 et ci et P_g dépend de la structure de O entre c_i et ses feuilles.

Nous avons expérimenté nos propositions et nous les avons analysées afin d'évaluer la part d'information extraite du corpus et celle extraite de la structure taxonomique. Nous avons également évalué leurs résultats par le biais de diverses comparaisons avec le jugement humain. Nous avons utilisé WordNet et des jeux de tests connus (de Rubsenstein& Goodenough [1965], de Miller&charles[1991] et de Finkelstein *et al.* [2002]) et le British National Corpus [Pedersen *et al.* 2004] pour le calcul du contenu informationnel en utilisant la formule de Resnik.

Nous avons tout d'abord déterminé la corrélation du contenu informationnel défini avec la formule de Resnik et un corpus (P_r) et celui déterminé par chacune de nos approximations sans utiliser un corpus. Les résultats ont montré une forte corrélation entre les valeurs du contenu informationnel déterminé avec P_r et P_g. De manière prévisible, le contenu informationnel basé sur P_p est le moins corrélé avec celui qui exploite le corpus. De plus, l'étude de la corrélation avec le jugement humain a montré que les corrélations obtenues avec P_r et P_g sont très proches et globalement les écarts sont négligeables [Blanchard, 2008].

Nous pouvons conclure à partir de ces expérimentations que la corrélation entre P_g et P_r montre que la masse d'information extraite du corpus en plus de celle inhérente à la structure taxonomique est relativement restreinte. Cependant, les résultats de ces expérimentations dépendent de l'ontologie et du corpus utilisés et ils ne peuvent être généralisés qu'avec des travaux portant sur d'autres ontologies et d'autres corpus.

Par ailleurs, nous avons montré dans Blanchard *et al.* [2008b] que plusieurs mesures de la litérature peuvent être reformulées en instanciant la forme unifiée proposée dans la section 4.3 avec une de nos approximations. Par exemple la mesure de similarité Wu & Palmer [1994] peut être reformulée en utilisant la formule (4.3) avec $\beta=\alpha=1/2$ et l'approximation de IC avec P_p ; la similarité de Stojanovic peut être réécrite avec la formule (4.3) avec $\beta=\alpha=1$ et l'approximation de IC avec P_p ; la similarité proposée dans [Blanchard *et al.* 2006b] coincide avec la formule (4.3) avec $\beta=\alpha=1/2$ et une approximation de IC avec P_s. Nous avons montré aussi que la redéfinition de IC par Seco *et al.* [2004] correspond à une normalisation de IC issue de P_s et la première approximation de IC par Sanchez *et al.* [2012] correspond (à 1 prêt) à l'approximation de IC avec P_g.

4.5 Mesures sémantiques de comparaison de deux ensembles de concepts

Certaines mesures pour la comparaison de deux ensembles de concepts sont une extension d'une mesure de comparaison de deux concepts. La façon la plus simple pour définir ce type d'extension est de définir une fonction qui permet l'agrégation des résultats des mesures sémantiques pour deux concepts (*e.g.* une somme, une moyenne). Par exemple, l'extension de la mesure sémantique de Rada *et al* [1989] à une mesure de comparaison de deux ensembles, le premier qui comprend k concepts C_i et le deuxième m concepts C'_j, est définie par [Rada *et al.* 1989] :

$$\text{Distance}(\{C_1, C_2....C_k\}, \{C'_1, C'_2....C'_m\}) = \frac{1}{km}\sum_{i=1}^{k}\sum_{j=m}^{m}\text{Distance}(C_i, C'_j) \quad (4.12)$$

Un autre point de vue considère seulement la meilleure valeur de la mesure sémantique entre une paire de concepts de deux ensembles. Dans ce cas, une mesure asymétrique exprime la contribution sémantique des concepts d'un ensemble en relation aux concepts de l'autre ensemble. Elle peut être définie comme suit [Azuaje *et al.* 05] :

$$\text{Distance}(\{C_1, C_2...C_k\}, \{C'_1, C'_2...C'_m\}) = \frac{1}{k+m}(\sum_{i=1}^{k}\min_j(\text{Distance}(C_i, C'_j)) + \sum_{j=1}^{m}\min_i(\text{Distance}(C_i, C'_j))) \quad (4.13)$$

Ces deux mesures ne sont pas normalisées et la deuxième n'est pas symétrique. Ces formes d'agrégation de distances entre deux concepts ont été utilisées aussi pour agréger des similarités entre deux ensembles de concepts dans [Mazandu&Mulder, 2013b].

La distance de Hausdorff peut être appliquée pour comparer deux ensembles de concepts. Elle est basée sur la détermination du maximum de la distance minimum d'un concept d'un ensemble avec les concepts de l'autre ensemble. Cette distance est normalisée mais elle se base sur la distance entre une seule paire de concepts [Euzenat, 2008]. La distance de couplage maximal de poids minimal est basée sur la maximisation du nombre de couples de concepts tels que la somme des distances de ces couples soit minimale [Valtechev, 1999].

D'autres mesures sémantiques pour la comparaison de deux ensembles de concepts utilisant la structure taxonomique et une extension ont été proposées. Elles sont basées sur le principe de l'utilisation de la quantité d'information commune apportée par les deux ensembles de concepts et la quantité d'information qui les décrivent [Pesquita *et al.* 2008], [Mazandu&Mulder, 2012], [Mazandu&Mulder, 2013b]. Par exemple, Pesquita *et al.* [2008] ont proposé la formule 4.14 pour le calcul de la similarité de deux protéines. Chaque protéine p ou q est annotée par un ensemble de concepts de l'ontologie OG, A^x_p est l'ensemble des concepts qui annotent p et leurs ancêtres.

$$SimGIC(p,q) = \frac{\sum_{x \in A_p^x \cap A_q^x} IC(x)}{\sum_{x \in A_p^x \cup A_q^x} IC(x)} \qquad (4.14)$$

Mazandu et Mudler [2012 ; 2013b] ont proposé d'autres formes pour le calcul de la similarité de deux protéines annotées chacune par un ensemble de concepts (4.15 et 4.16).

$$SimDIC(p,q) = \frac{2X \sum_{x \in A_p^x \cap A_q^x} IC(x)}{\sum_{x \in A_p^x} IC(x) + \sum_{x \in A_q^x} IC(x)} \qquad (4.15)$$

$$SimUIC(p,q) = \frac{\sum_{x \in A_p^x \cap A_q^x} IC(x)}{\max(\sum_{x \in A_p^x} IC(x), \sum_{x \in A_q^x} IC(x))} \qquad (4.16)$$

Dans ces approches, le calcul du contenu informationnel commun de deux ensembles de concepts est effectué en déterminant tout d'abord l'ensemble des ancêtres en commun des deux ensembles de concepts, et en additionnant ensuite les contenus informationnels des concepts de cet ensemble. Cette méthode de calcul ne respecte pas une des propriétés du contenu informationnel à savoir que l'information apportée par un ancêtre est déjà apportée par son fils (cf. section 4.3). D'ailleurs, si on utilise les formules (4.14), (4.15) ou (4.16) pour comparer deux protéines annotées chacune par un seul concept, le contenu informationnel commun de deux protéines sera équivalent au contenu informationnel commun de ces deux concepts et il sera calculé par la somme des contenus informationnels des ancêtres de leur subsumant commun le plus spécifique, y compris ce dernier. Cependant, la majorité des travaux dans la littérature définit le contenu informationnel de deux concepts (dans une ontologie sans multi-héritage[35]) par le contenu informationnel de leur subsumant commun le plus spécifique.

En plus, dans ces approches, la forme de la formule du dénominateur ne représente pas toujours la description du contenu informationnel de deux ensembles. Par exemple, dans la formule (4.16) le dénominateur représente le maximum des descriptions des deux ensembles, ce qui implique qu'il ne représente qu'une seule des deux descriptions. Par conséquent, il est difficile d'adapter ces mesures à des objectifs différents de comparaison (*e.g.* équivalence, inclusion).

Dans nos travaux, nous avons réutilisé la forme unifiée des mesures sémantiques pour la comparaison de deux concepts (4.2) et nous l'avons étendue pour la définition des mesures sémantiques Ms pour la comparaison de deux ensembles de concepts Si et Sj [Blanchard, 2008], [Blanchard *et al.* 2008a] :

[35] Pour le calcul du contenu informationnel dans le cas d'une ontologie avec un multi-héritage, le lecteur peut se référer au chapitre 5 de la thèse d'E. Blanchard.

$$Ms(S1, S2) = \frac{ICa^{\cap}(Si,Sj)}{\alpha ICa^{-}(Si, Sj) + \beta ICa^{-}(Si,Sj) + ICa^{\cap}(Si,Sj)} \qquad (4.17)$$

où $ICa^{\cap}(Si,Sj)$ est une approximation du contenu informationnel commun à deux ensembles de concepts et $ICa^{-}(Si,Sj)$ (respectivement $ICa^{-}(Si,Sj)$) est une approximation du contenu informationnel distinctive de Si par rapport Sj (respectivement Sj par rapport à Si).

Soit Sub(Si) l'ensemble de tous les subsumants des concepts de Si (y compris les concepts de Si eux-mêmes) :

$$Sub(Si) = \{c / ci \in Si \wedge ci \sqsubseteq c\} \qquad (4.18)$$

L'approximation du contenu informationnel d'un ensemble de concepts ICa(Si) est définie par la somme du contenu informationnel de chaque concept de Sub(Si) moins le contenu informationnel de son subsumant direct[36] :

$$ICa(Si) = ICa(c0) + \sum_{C \in Sub(Si)} ICa(c) - ICa(Père(c)) \qquad (4.19)$$

où c0 est la racine de l'ontologie considérée.

Les deux définitions suivantes, analogues aux définitions (4.4) et (4.5) pour les mesures sémantiques de comparaison de deux concepts, permettent de compléter la définition de (4.17) :

$$ICa^{\cap}(S_i, S_j) = ICa(Sub(S_i) \cap Sub(S_j)) \qquad (4.20)$$

$$ICa^{-}(S_i, S_j) = ICa(S_i) - ICa^{\cap}(S_i, S_j) \qquad (4.21)$$

En appliquant ces deux formules pour la comparaison des CT1, CT2 et CT3 de l'exemple de la section 4.2, en utilisant l'annotation avec un ensemble de concepts, nous obtenons :

Sub(As(CT1))={C9, C8, C5, C3, C0, P6, P4, P2, P0}
Sub(As(CT2))={C9, C8, C5, C3, C0, P6, P4, P2, P0}
Sub(As(CT3))={C9, C8, C5, C3, C0, P7, P6, P4, P2, P0}
$ICa^{\cap}$(As(CT1), As(CT2)) = ICa(CT1)= ICa(CT2),
ICa^{-}(As(CT1),As(CT2)) = 0
$ICa^{\cap}$(As(CT1), As(CT3)) = ICa({C9, C8, C5, C3, C0, P6, P4, P2, P0})
ICa^{-}(As(CT1),As(CT3))= 0 , ICa^{-}(As(CT3, As(CT1))= ICa(P7)- ICa(P4)
$ICa^{\cap}$(As(CT2), As(CT3)) = ICa({C9, C8, C5, C3, C0, P6, P4, P2, P0}).
ICa^{-}(As(CT2),As(CT3))= 0 , ICa^{-}(As(CT3, As(CT2))= ICa(P7)- ICa(P4)

Nous avons expérimenté cette famille de mesures dans le cadre de l'approche UEML pour la comparaison de constructs de langages de modélisation [Blanchard, 2008], [Blanchard *et al.* 2008a], [Anaya *et al.* 2010] (cf. section 4.7). Nous l'avons utilisée également pour l'identification sémantique des communautés dans un réseau social où nous avons comparé les centres d'intérêt d'une communauté avec les centres d'intérêt d'une autre communauté ou d'un individu [BenAmor *et al.* 2016]. Nous ne l'avons pas comparée aux autres mesures présentées dans cette section. Toutefois, notre proposition est théoriquement plus intéressante que les autres. Tout

[36] Cette formule reste la même dans le cas d'une ontologie avec un multi-héritage (cf. [Blanchard, 2008])

d'abord, la forme de cette famille de mesures représente bien le principe de comparaison de deux objets en théorie de l'information. En plus, la définition du contenu informationnel d'un ensemble de concepts respecte bien le principe que l'information apportée par un concept est déjà apportée par son fils, ce qui n'est pas toujours le cas dans les autres mesures proposées.

4.6 Mesures sémantiques de comparaison de deux graphes sémantiques

Comme nous avons vu dans la section 4.2, un objet peut être annoté par un graphe sémantique. Ce type de graphe est particulier car il peut comprendre plusieurs nœuds ou arcs étiquetés respectivement avec le label du même concept ou de la même relation.

La comparaison de graphes est une problématique étudiée dans plusieurs domaines [Darmont&Bossard, 2006]. Certains travaux s'intéressent aux graphes de taille importante et cherchent à évaluer la complexité de leur structure. Cette complexité est habituellement évaluée par une mesure du contenu informationnel d'un graphe, appelé mesure de l'entropie de graphe [Mowshowitz&Mitsou, 2009], [Dehmer&Mowshowitz, 2011]. D'autres travaux portant particulièrement sur les petits graphes s'intéressent à les comparer : (1) en faisant de l'appariement de graphes ou (2) en évaluant la distance d'édition entre eux [Bunke, 1997].

Nous nous sommes intéressés à ces derniers travaux parce que nous considérons des petits graphes sémantiques et non la complexité de la structure des grands graphes. Nous avons proposé une famille de mesures sémantiques pour la comparaison de deux objets annotés par deux graphes sémantiques. Cette famille de mesures a la même forme unifiée que celle proposée pour les deux familles de mesures précédentes et elle est définie suivant la même approche.

L'approche proposée dans [Bunke&Shearer, 1998], [Riesen&Bunke, 2010] pour la comparaison de deux graphes étiquetés utilise le graphe commun maximum de ces graphes. Cependant, pour déterminer ce type de graphe, elle effectue un appariement exacte de ces deux graphes ce qui ne permet pas de déterminer leur proximité sémantique s'ils sont étiquetés par des concepts différents mais proches sémantiquement. La distance sémantique de deux graphes par la méthode d'édition requiert d'effectuer tous les appariements possibles entre deux graphes et ensuite de mesurer la distance entre chaque deux nœuds appariés, par l'effort nécessaire pour passer de l'étiquette (i.e. le label) d'un nœud à celle de l'autre [Bunke, 1997]. La distance entre deux graphes est calculée par la somme des distances d'édition entre les nœuds appariés divisée par le nombre de paires de nœuds appariés. La distance déterminée par cette méthode est normalisée et prend en compte un appariement approximé sémantiquement (un nœud avec une étiquette donnée peut être apparié avec un nœud ayant une étiquette différente) mais elle n'est pas basée sur le principe de comparaison d'objets en théorie de l'information. En plus, si le coût d'une opération de transformation d'une étiquette d'un nœud en une autre étiquette peut être évalué par une distance sémantique entre étiquettes, le coût d'une opération de rajout d'un nouvel nœud est difficile à déterminer d'une façon homogène à celle permettant de déterminer le premier type de coût.

4.6.1 Forme unifiée des mesures de comparaison de graphes sémantiques

Par analogie avec la comparaison de deux concepts ou de deux ensembles de concepts, nous avons utilisé la même forme unifiée (cf. 4.2 et 4.17) pour définir la famille de mesures sémantiques Mg pour comparer deux graphes sémantiques G1 et G2 :

$$Mg(G1,G2) = \frac{ICa^{\cap}(G1,G2)}{\alpha ICa^{-}(G1,G2) + \beta ICa^{-}(G1,G2) + ICa^{\cap}(G1,G2)} \qquad (4.22)$$

Pour définir $ICa^{\cap}$ (G1, G2), nous avons introduit la notion de « graphe sémantique commun maximum » (maximal semantic common graph) (mscg) G de deux graphes G1 et G2 défini par un *graphe sémantiquement commun à G1 et G2* ayant un contenu informationnel maximum :

$$G \text{ est un mscg}(G1, G2) \text{ si } G \in \{G_{mk} / mk \in M\} \text{ et } IC(G) = \max_{mk \in M} IC(G_{mk}) \qquad (4.23)$$

où M est l'ensemble de tous les cas possibles d'appariement sémantique de G1 et G2. Ce type d'appariement est défini dans la section suivante.

Comme pour la comparaison de deux concepts ou de deux ensembles de concepts, la définition (4.22) est complétée par :

$$ICa^{\cap}(G1, G2) = ICa(G) \qquad (4.24)$$

$$ICa^{-}(G1,G2) = ICa(G1) - ICa^{\cap}(G1, G2) = ICa(G1) - ICa(G). \qquad (4.25)$$

Dans ces définitions, nous avons utilisé comme pour la comparaison de deux concepts ou de deux ensembles de concepts la notion du contenu informationnel d'un graphe sémantique. Pour respecter la définition originelle du contenu informationnel d'un concept, la définition du contenu informationnel d'un graphe sémantique doit être proportionnelle aux nombres de ses nœuds et ses arcs. En effet, plus il y a de nœuds et d'arcs dans un graphe, moins est le nombre d'instances pouvant être associées à chacun d'entre eux. Cependant, comme dans le cas de concepts et d'ensembles de concepts, le contenu informationnel ne peut être évalué que d'une façon approximative.

4.6.2 Appariement de graphes sémantiques et graphe sémantique commun

Nous appliquons sur deux graphes sémantiques un appariement inexact et approché. Un appariement est inexact car chaque nœud du premier graphe (respectivement du deuxième graphe) n'est pas nécessairement apparié à un nœud du deuxième graphe. Il est approché car un nœud avec un label donné peut être apparié à un nœud avec un label différent. Nous définissons le résultat de ce type d'appariement par un graphe sémantique commun. Chaque nœud de ce graphe correspond à un couple de nœuds appariés ayant le label du subsumant commun le plus spécifique de deux concepts de ce couple. Nous définissons dans la suite la notion de graphe sémantique commun et l'appariement permettant de déterminer ce graphe.

Un graphe sémantique G est défini par $G=(V, E, \alpha_v, \alpha_e, O)$ où V et E sont respectivement les nœuds et les arcs de G et $E \subseteq V \times V$, O est une ontologie, (C(O) est l'ensemble de concepts de O, R(O) est l'ensemble de ses relations), α_v: $V \rightarrow C(O)$ est une fonction qui associe un nœud à un concept, α_e: $E \rightarrow R(O)$ est une fonction qui associe un arc à une relation (y compris les arcs et relations *inférés*) telle que si $e_{ij}=(v_i, v_j)$ alors $\alpha_e(e_{ij}) \subseteq \alpha_v(v_i) \times \alpha_v(v_j)$ dans O.

Soient deux graphes sémantiques $G1=(V1, E1, \alpha_{v1}, \alpha_{e1}, O)$ et $G2=(V2, E2, \alpha_{v2}, \alpha_{e2}, O)$ et M l'ensemble de tous les appariements inexacts et approchés possibles de G1 et G2.

Nous avons défini un *appariement inexact et approché* m_k de M par deux fonctions mv_k et me_k de mappings partiels (fonctions injectives) et un à un (one-to-one). Cette définition et ses contraintes garantissent que tous les appariements possibles de G1 avec G2 sont les mêmes que

ceux de G2 avec G1. Nous avons ensuite défini un « graphe sémantique commun », à partir d'un mapping mk∈ M par un graphe G_{mk} = (V, E, α_v, α_e, O). G_{mk} est composé de nœuds tels que chacun représente un couple de nœuds de G1XG2 explicitement apparié par mk. Chaque nœud de G_{mk} a le label du concept qui est le subsumant commun le plus spécifique des concepts ayant les labels de son couple de nœuds (cf. la définition formelle de G_{mk} dans [Harzallah&Berio, 2015]).

Fig. 4.5 illustre un graphe sémantique commun des deux graphes de représentation ontologique des constructs « Activity_Edge » d'UML et « Parallel_Gate » de BPMN, en utilisant les taxonomies Class et Property d'UEMO (cf. Fig. 2.2 et Fig. 2.3 du chapitre 2).

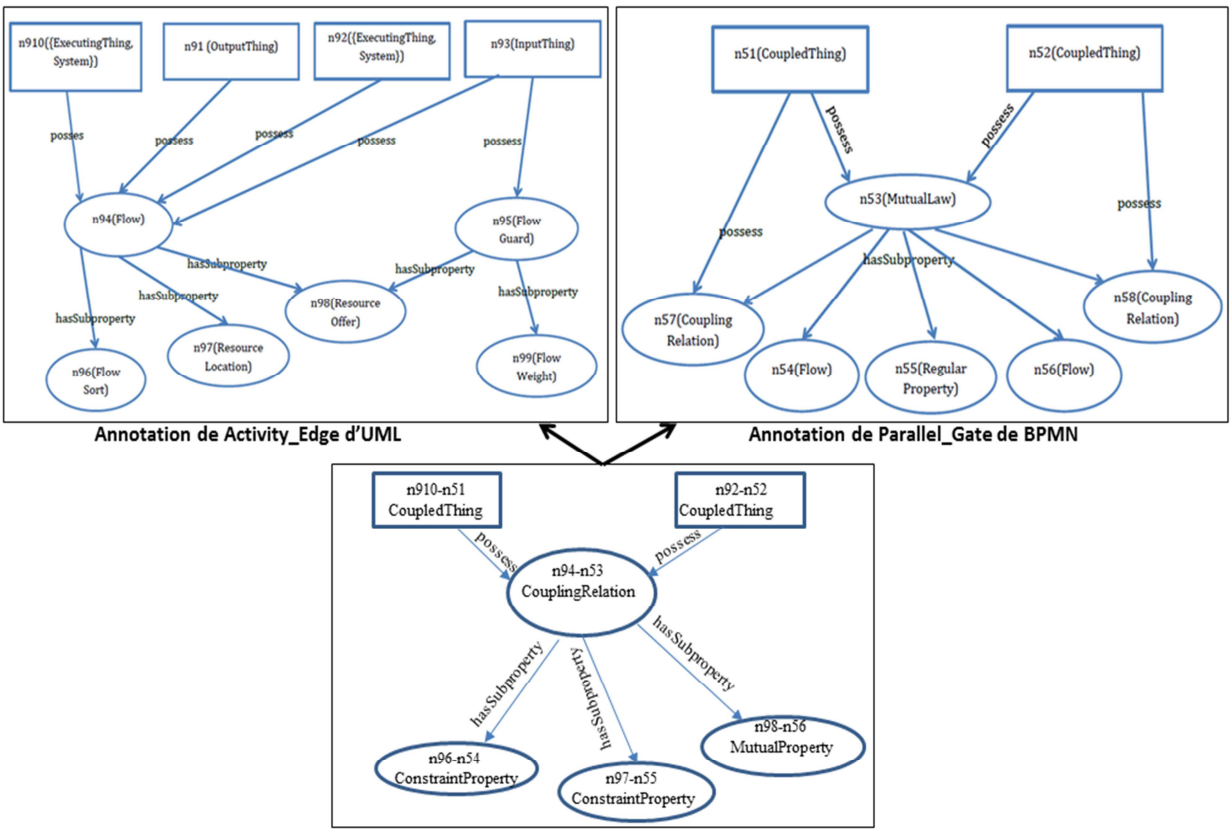

Fig. 4.5 – Graphe sémantique commun des graphes d'annotation de Activity_Edge et Parallel Gate

G_{mk} est isomorphe à un sous-graphe sémantique de G1 et de G2. La définition de G_{mk} est plus contraignante que la description commune d'un sous-graphe dans [Champin&Solnon, 2003] : deux nœuds de G1 liés par un arc ne peuvent être alignés qu'à deux nœuds de G2 liés par un arc ayant le même label que celui qui relie les deux premiers nœuds. En effet, quand l'annotation d'un objet est faite par un graphe, nous considérons que les relations de ce graphe sont aussi importantes que ses concepts (sinon, on utilisera une annotation avec un ensemble de concepts).

4.6.3 Approximation du contenu informationnel d'un graphe sémantique

Nous avons proposé une approximation du contenu informationnel d'un graphe sémantique basée sur la proposition de Champin&Solon [2003], mais en remplaçant le poids associé à un nœud, défini d'une façon subjective par l'utilisateur, par une approximation de son contenu

informationnel. Les arcs (ou relations) inférés ne sont pas pris en compte dans cette approximation afin d'éviter de considérer l'information apportée par un arc plusieurs fois.

Notre approximation du contenu informationnel d'un graphe sémantique G, notée ICa(G), est définie par :

$$ICa(G) = \sum_{v_i \in V} ICa(v_i) + \sum_{(v_i, v_j) \in E(G)} \frac{ICa(v_i) + ICa(v_j)}{2} \quad (4.26)$$

Dans cette formule (4.26), le contenu informationnel d'un graphe augmente bien avec le nombre de ses nœuds et ses arcs. Cette formule peut être interprétée d'un autre point de vue. Dans un graphe sémantique, les arcs représentent des relations sémantiques qui connectent des nœuds représentés par des concepts. Les relations sémantiques contribuent souvent à la définition des concepts. La conséquence est que le contenu informationnel d'un concept augmente avec le nombre de ses relations.

La formule 4.27 met en évidence la partie de la formule 4.26[37] associée à chaque nœud.

$$ICa(G) = \sum_{v_i \in V} (ICa(v_i) + \sum_{(v_i, v_j) \in E} \frac{ICa(v_i) + ICa(v_j)}{4}) \quad (4.27)$$

où

$$ICa(v_i) + \sum_{(v_i, v_j) \in E} \frac{ICa(v_i) + ICa(v_j)}{4} = ICa(v_i) + |arcs(v_i)|(\frac{ICa(v_i)}{4} + \sum_{(v_i, v_j) \in E} \frac{ICa(v_j)}{4|arcs(v_i)|}) \quad (4.28)$$

|arcs(vi)| étant le nombre d'arcs liés à vi.

La formule 4.28 est intéressante car elle peut être considérée comme la définition du contenu informationnel d'un nœud v_i en prenant en compte les relations auxquelles il est connecté. En effet, elle correspond à ce que dans [Pirrò&Euzenat, 2010] est appelé le contenu informationnel étendu d'un concept (« *the extended information content* of concept »), noté eIC(c), et défini par eIC(c)=ζiIC(c)+ηEIC(c), où iIC(c) est une approximation de IC(c) dans une taxonomie et EIC(c) est une moyenne des contenus informationnels des concepts reliés à c par des relations autres que les relations taxonomiques, ζ et η sont deux paramètres. eIC(c) est proche (quasi-similaire) de la partie suivante de la formule (4.28) quand on fixe ζ= η=1/4 :

$$\frac{ICa(v_i)}{4} + \sum_{(v_i, v_j) \in E} \frac{ICa(v_j)}{4|arcs(v_i)|} \quad (4.29)$$

Par conséquent, notre définition d'IC(G) intègre le contenu informationnel étendu d'un concept dans une ontologie.

4.7 Application des trois familles de mesures dans l'approche UEML

L'application de la famille de mesures de comparaison de graphes sémantiques à la comparaison de CT1, CT2 et CT3 (cf. section 4.2), en utilisant l'annotation avec un graphe sémantique donne :

[37] Les formules 4.26 et 4.27 sont équivalentes.

ICa(G1)=ICa(C9)+ICa(P6)+ ICa(P4) +(ICa(C9)+ICa(P6))/2+(ICa(C9)+ICa(P4))/2
ICa(G2) = ICa(C9)+ICa(P6) +(ICa(C9)+ICa(P6))/2
ICa(G3) = ICa(C9)+ICa(P6)+ ICa(P7) +(ICa(C9)+ICa(P6))/2+(ICa(C9)+IC(P7))/2.

Nous n'illustrons ici que le calcul de $ICa^{\cap}$(G1, G2). Il n'y a que deux appariements possibles de G1 et G2 qui concernent plus qu'un concept de G1.

mv1(n11)= n21, mv1(n12)= n22, me1((n11, n12, possesses)) = (n21, n22, possesses)
ICa(Gm1) = ICa(msc(c9, c9))+ICa(msc(p6, p6)) + ((ICa(msc(c9, c9))+ IC(masc(p6, p6)))/2
mv2(n11)= n21, mv2(n13)= n22, me2((n11, n13, possesses)) = (n21, n22, possesses)
IC(Gm2)= IC(C9)+IC(C4)+(IC(C9)+IC(C4))/2
Parce que IC(Gm1)> IC(Gm2), Gm1 est le graphe sémantique commun maximum
$ICa^{\cap}$(G1, G2)= IC(Gm1), ICa^{-}(G1, G2)) = IC(P4) +(IC(C9)+IC(P4))/2, ICa^{-}(G2, G1)) = 0.

Pour comparer les résultats des trois familles de mesures en utilisant ce même exemple, nous avons appliqué à celui-ci trois similarités, chacune appartenant à une de ces trois familles, en instanciant la forme unifiée avec β=α=1. Tab. 4.1 comprend les résultats des trois similarités. La similarité utilisant une annotation avec un seul concept donne des égalités pour toutes les paires de constructs parce les « phénomènes centraux » des trois constructs sont appariés chacun au même concept d'UEMO. Avec cette mesure, ces constructs ne peuvent pas être distingués. La similarité utilisant une annotation avec un ensemble de concepts ne permet pas de distinguer CT1 et CT2, mais elle distingue bien CT1 et CT3, et CT2 et CT3. Enfin, la troisième mesure, comme attendu, distingue les trois constructs, en déterminant que CT3 est plus proche sémantiquement de CT1 que de CT2.

Paire de constructs	Annotation avec un seul concept	Annotation avec un ensemble de concepts	Annotation avec un graph sémantique
Sim(CT1,CT2)	1	IC(C9)+IC(P6)/(IC(C9) +IC(P6))=1	IC(C9)+IC(P6)/((4/3IC(C9)+IC(P4)+IC(P6))
Sim(CT1,CT3)	1	(IC(C9)+IC(P6))/(IC(C9)+ IC(P6)+IC(P7)-IC(P4))	4/3IC(C9)+IC(P6)+IC(P4))/(4/9IC(C9)+IC(P6)+IC(P7)
Sim(CT2,CT3)	1	(IC(C9)+IC(P6))/(IC(C9)+IC(P6)+IC(P7)-IC(P4))	IC(C9)+IC(P6)/4/3IC(9)+IC(P6)+IC(P7)

Tab. 4.1 - Résultats des trois similarités pour CT1, CT2 et CT3

Nous avons expérimenté ces trois familles de mesures sémantiques de comparaison sur des constructs incorporés dans l'approche UEML. La troisième famille de mesures est basée sur un appariement de graphes sémantiques. Ce dernier est un problème NPdifficiles. Certaines approches pour l'appariement de graphes utilisent des algorithmes incomplets qui ne garantissent pas l'optimalité de la solution trouvée mais ils ont une complexité polynomiale [Boeres *et al.* 2004]. Comme dans d'autres approches [Champin&Solnon, 2003], [Ambauen *et al.* 2003], nous avons défini un algorithme complet pour la recherche de l'appariement qui maximise le contenu informationnel du graphe sémantique commun. Notre algorithme est basé sur une exploration exhaustive de l'espace de recherche avec des techniques de filtrage. Toutefois, nous nous sommes limités à la comparaison des très petits graphes (huit nœuds maximum ayant comme label celui d'un concept d'une même taxonomie), afin d'éviter une explosion combinatoire.

Dans cette expérimentation, nous avons instancié la forme unifiée pour ces trois familles avec β=α=1 et nous l'avons appliquée aux 46 pairs de constructs formés avec dix constructs : IDEF3-SPL, IDEF3-AND, IDEF3-RCD-XOR, ARIS_And, BPMN_ParallelGate, UML_Action,

IDEF3_UOB, UML_Activity, UML_ActivityEdge et IDEF3_RCD_OR de quatre langages : IDEF3, ARIS, BPMN et UML. L'incorporation de ces constructs dans l'approche UEML a été déjà réalisée dans le Rex INTEROP. Nous avons ensuite comparé les résultats de ces trois mesures aux valeurs du jugement humain. Ces dernières ont été établies par cinq experts qui ont participé à l'incorporation des langages dans l'approche UEML : Andreas Opdahl (professeur à l'université de Bergan), Mickeal Petit (Professeur à l'université de Namur), Giuseppe Berio (Prodesseur à l'université de Bretagne Sud), Hervé Panetto (professeur à l'université de Lorraine) et moi-même.

L'analyse préliminaire des résultats a montré des oublis et des erreurs dans l'annotation des constructs qui ont affecté les résultats de la mesure. Pour cela, nous avons supprimé deux constructs : IDEF3-RCD-XOR et IDEF3_RCD_OR car leur annotation était erronée. L'analyse des valeurs de similarité des huit constructs restants a montré que la mesure de comparaison en utilisant une annotation avec un graphe sémantique a un écart moyen par rapport au jugement humain moins important que celui des deux autres mesures. En revanche, la mesure de comparaison utilisant une annotation avec un seul concept a une capacité discriminante élevée quand il s'agit de comparer deux constructs ayant des phénomènes centraux de classes différentes.

Par ailleurs, nous avons remarqué que les deux dernières mesures donnent parfois une forte similarité entre deux constructs pour lesquels on doit avoir une similarité égale à zéro, comme par exemple pour les constructs UML_Activity et BPMN_ParallelGate. L'étude de ce cas, nous a permis d'identifier l'intérêt de prendre en compte le phénomène central d'un construct dans les deux dernières mesures afin d'améliorer leur résultats. En effet, on peut avoir deux constructs annotés par quasiment les mêmes concepts et ayant une structure syntaxique proche, mais ils représentent des phénomènes complétement différents. L'un des deux représente une propriété d'une classe et le deuxième représente cette classe avec ses propriétés.

4.8 Analyse des mesures sémantiques de comparaison de deux concepts

Le choix d'une mesure est une problématique qui a été évoquée depuis longtemps dans la littérature et à plusieurs occasions. Elle consiste à déterminer la mesure qui donne les résultats les plus précises pour une application donnée. La précision d'une mesure est souvent déterminée en fonction de la proximité de ses résultats au jugement humain pour une application et un contexte donnés. Des travaux ont développé des outils sur le web pour comparer des mesures de similarité proposées dans la littérature et plus particulièrement pour le domaine bioinformatique (*e.g.* http://neurolex.org/wiki/Category:Resource:Gene_Ontology_Tools). Ces outils aident à l'exploration de ces mesures et le choix d'une d'entre elle [Du *et al.* 2009], [Caniza *et al.* 2014], [Harispe *et al.* 2014]. Toutefois, si on change de contexte ou de jeu de données la précision d'une même mesure peut changer [Gaston *et al.* 2014].

Certains travaux ont abordé cette problématique en étudiant des corrélations entre les résultats des mesures ou en déterminant des liens théoriques entre leur formule [Cha *et al.* 2010], [Cross *et al.* 2013]. Harispe *et al.* [2014] se sont intéressés au choix d'une mesure par rapport aux valeurs de α et β et donc par rapport à la prise en compte de l'information commune à deux concepts et l'information distinctive de chacun par rapport à l'autre. Ceci revient à choisir les paramètres α et β d'une mesure en fonction de son objectif : similarité, inclusion, *etc.*

Dans Blanchard *et al.* [2008], nous avons identifié un lien entre la famille des mesures de similarité σ_β (β étant un réel strictement positif) appliquées à deux ensembles de données A et B

et définie par (4.31) et la forme unifiée des mesures sémantiques pour la comparaison de deux concepts (4.3).

$$\sigma_\beta(|A|,|B|,|A \cap B|) = \frac{\beta.|A \cap B|}{|A| + |B| + (\beta - 2)|A \cap B|} \qquad (4.31)$$

En effet, en remplaçant, un ensemble par un concept et la cardinalité d'un ensemble de données par une approximation du contenu informationnel d'un concept on obtient une famille de mesures sémantiques respectant la forme unifiée avec α=β ≠ 0. Cette dernière a les mêmes caractéristiques que la famille de mesures de similarité σ_β.

$$\sigma_\beta{}'(c_i,c_j) = f_\beta(IC(c_i),IC(c_j),IC^\cap(c_i,c_j)) = \frac{\beta.IC^\cap(c_i,c_j)}{IC(c_i) + IC(c_j) + (\beta - 2).IC^\cap(c_i,c_j)} \qquad (4.32)$$

Or les mesures de la famille σ_β ont la même préordonnance[38] (i.e. elles suivent le même ordre) ce qui implique que les mesures sémantiques ayant la forme unifiée et instanciées en variant α et β, telles que α= β ≠0, suivent le même ordre [Blanchard *et al.* 2008b]. Une mesure de cette famille est une similarité qui se différencie d'une autre mesure de cette même famille par l'importance qu'on souhaite donner à la partie commune de deux concepts à comparer.

Par ailleurs, le choix d'une mesure peut dépendre d'autres paramètres qui ne sont pas abordés dans la littérature. Selon la forme unifiée d'une mesure sémantique, les résultats d'une mesure sémantique dépendent : (1) des valeurs de α et β, (2) de l'approximation de IC, et (3) de l'ontologie utilisée. Quid du comportement d'une mesure si on change l'approximation du contenu informationnel ou on change l'ontologie sur laquelle elle est appliquée ?

4.8.1 Hypothèse inhérente à une mesure sémantique

Le contenu informationnel peut être approximé en utilisant plusieurs méthodes et plusieurs types de données en entrée. Toutefois, la diversité et la multiplication des propositions rendent difficile à en sélectionner une. Lee *et al.* [2008] trouvent que les mesures utilisant seulement la structure taxonomique (*e.g.* profondeur d'un concept, nombre de relations taxonomiques entre deux concepts) satisfont mieux les experts et elles sont mieux que celles utilisant la structure taxonomique et une extension (*e.g.* des instances ou des occurrences de concepts). Zavitsanos *et al.* [2010] ont montré que les mesures qui utilisent la structure taxonomique et une extension donnent des meilleurs résultats. Nous avons montré dans nos travaux que les mesures qui utilisent l'approximation ICg, utilisant seulement la structure taxonomique, donnent des résultats proches de celles utilisant la structure taxonomique et un corpus de textes (cf. section 4.6.3). Que se passe-t-il ? En réalité, il y a des caractéristiques intrinsèques des mesures qui les rendent plus ou moins adaptées au contexte d'une application donnée. En effet, une mesure est basée sur une ou plusieurs hypothèses implicites ou explicites portant sur l'ontologie utilisée. Par exemple, considérons le type de mesure qui utilise la structure taxonomique et qui considère que la similarité de deux concepts dépend du nombre de relations taxonomiques entre eux. Ce type de mesures implique que toutes les relations *est-un* dans une ontologie représentent, d'une façon approximative, le même degré de spécialisation Père-Fils. La qualité des résultats de ce type de mesures va dépendre, en partie du niveau de respect de cette hypothèse dans l'ontologie utilisée. Par exemple, dans l'extrait d'une taxonomie du domaine des moyens de transports (Fig. 4.6) « Moyen de transport aérien » *est-un* « Moyen de transport » et « vélo » *est-un* « Moyen de Transport ». Il est clair que ces deux relations ne représentent pas le même degré de spécialisation

[38] Deux mesures M1 et M2 ont la même préordonnance si quelques soient les concepts C1, C2, C3 et C4, M1(C1,C2)≤ M1(C3,C4) <=> M2·(C1,C2) ≤ M2·(C3,C4).

Père-Fils. Une mesure basée sur cette hypothèse et appliquée à cet extrait donne que « Moyen de Transport » est aussi similaire à « Moyen de transport aérien » qu'à « Vélo » alors que d'une façon intuitive, il est normal de dire que « Moyen de Transport » est plus similaire à « Moyen de transport aérien » qu'à « Vélo ». L'hypothèse de cette mesure n'est pas respectée par cette taxonomie, cette mesure appliquée à cette taxonomie ne donnera pas des résultats précis.

Il est clair que pour bien choisir une mesure, son hypothèse doit être tout d'abord bien explicite, comprise et acceptable. Sinon, comme mentionné dans [Zavitsanos *et al.* 2010], sans bien comprendre l'hypthèse d'une mesure, l'appliquer implique une faible précison de ses résultats. L'hypothèse d'une mesure peut aussi influencer sa robustesse [Mathur&Dinakarpandian, 2012]. Cette dernière quantifie la variation de ses résultats suite à un changement dans l'ontologie sur laquelle elle est appliquée. Par exemple, une mesure utilisant que la structure taxonomique est robuste à l'ajout ou à l'enlèvement des relations non-taxonomiques. Par contre, ajouter des nouveaux concepts intermédiaires qui subsument ou qui sont subsumés par d'autres concepts peut modifier les résultats d'une mesure ayant l'hypothèse que toutes les relations *est-un* representent le même degré de spécialisation Père-Fils.

Dans nos travaux, nous avons cherché à mettre en évidence l'hypothèse présumée sur les artefacts d'une ontologie pour chaque approximation faite pour le calcul du contenu informationnel.

Resnik [1995] a proposé d'approximer le contenu informationnel par la fréquence des occurrences d'un concept et de ses ancêtres (Pr) dans un corpus de textes. L'utilisation de l'approximation de Resnik demande la vérification de son hypothèse (Hy_O0) *i.e.* les concepts ayant le même nombre d'occurrences (leurs occurrences et ceux de leurs ancêtres) dans le corpus utilisé ont le même « degré de spécificité ». Si cette hypothèse n'est pas vérifiée, on ne peut pas prévoir une forte précision de la mesure qui l'utilise. Par exemple, si on considère un corpus portant sur les moyens de transport et l'extrait précédent (cf. Fig. 4.6). Si dans ce corpus la somme du nombre d'occurrences de « Vélo », « Vélo_non_électrique » et « Vélo_électrique » (les deux derniers sont des subsumés de « Vélo ») est égale à la somme des nombres d'occurrences de « Moyen_de_transport_aérien » et "Avion" (le dernier est subsumé par « Moyen_de_transport_aérien »), alors la mesure qui utilise l'approximation de Resnik pour l'estimation de IC donnera que « Moyen_de_transport » est aussi similaire à « Vélo » qu'à « Moyen_de_transport_aérien ». Evidemment, ce résultat n'est pas correct sémantiquement, mettant en cause la validité de l'hypothèse Hy_O0 pour cette taxonomie.

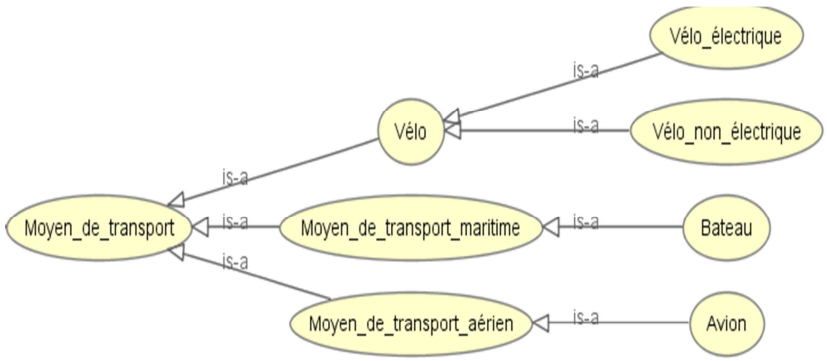

Fig. 4.6 - Extrait d'une taxonomie des moyens de transport

Nous avons proposé trois alternatives pour l'approximation d'IC utilisant que la structure d'une taxonomie [Blanchard, 2008], [Blanchard *et al.* 2008b] (cf. section 4.4). Ces approximations sont basées sur des hypothèses qui portent sur la distribution des instances d'une ontologie sur ses concepts. Pour la première approximation (P_p), les instances sont supposées être réparties d'une façon équivalente sur tous les concepts ayant la même profondeur dans une taxonomie. L'hypothèse ($H_{yO}1$) associée à cette approximation est « Toutes les spécialisations sont de même degré d'importance ». Pour la deuxième approximation (P_s), les instances d'un concept sont supposées réparties d'une façon égale sur ses fils, l'hypothèse ($H_{yO}2$) associée à cette approximation est « Le degré d'importance d'une spécialisation entre un concept et son fils dépend du nombre de fils de ce concept ». Pour la troisième approximation (Pg), les instances sont supposées être distribuées d'une façon égale sur toutes les feuilles d'une taxonomie, l'hypothèse (H3) associée à cette approximation est « Le degré d'importance d'une spécialisation entre un concept et un fils dépend de la différence du nombre de feuilles de ce concept et celui de ce fils ».

Nous avons montré précédemment que l'utilisation d'une mesure utilisant Hy_O1 et l'extrait de taxonomie de Fig. 4.6 donne des résultats non précis pour certains concepts de cet extrait.

L'utilisation d'une mesure utilisant $H_{yO}2$ et le même extrait donne que « Moyen_de_transport_aérien » est très similaire à « Avion » (parce que « Avion » est son unique fils). Ceci n'est pas vraiment correct. Enfin, si une mesure est basée sur $H_{yO}3$ et appliquée au même extrait de Fig. 4.6 elle donne que « Moyen_de_transport » est plus similaire à « Vélo » qu'à « Moyen_de transport _aérien » (parce que « Vélo » est lié à deux feuilles alors que « Moyen_de_transport_aérien » est lié à une seule feuille, par conséquent ICa(Vélo)<ICa(Moyen_de_ transport_aérien).

Avec ces exemples, nous ne revendiquons pas l'utilisation d'une mesure dont l'hypothèse est complètement respectée par l'ontologie sur laquelle elle est appliquée. Cependant, il est important de connaître l'hypothèse d'une mesure pour comprendre ses résultats une fois appliquée sur une ontologie afin de la choisir en connaissance de cause.

4.8.2 Analyse expérimentale des mesures en fonction de l'approximation d'IC.

Nous nous sommes intéressés particulièrement aux résultats des trois mesures de similarité S_p, S_s et S_g définies selon la forme unifiée avec $\beta=\alpha=1$ et ayant respectivement les hypothèses : $H_{yO}1$, $H_{yO}2$ et $H_{yO}3$. Nous avons constaté tout d'abord que ces mesures ne donnent pas nécessairement des résultats identiques. Nous distinguons deux principales différences : la première concerne la différence des valeurs des mesures et la deuxième concerne la différence de l'ordre de leurs valeurs. En plus, nous avons analysé les résultats d'une mesure avant et après un changement réalisé dans l'ontologie sur laquelle elle est appliquée et nous avons constaté une variation de ses résultats (valeurs et ordre des valeurs).

Nous illustrons ces deux constats à l'aide de l'exemple de Fig.4.7. Soient O et O' deux ontologies, O' est obtenue en rajoutant la feuille C10 à O (cf. Fig. 4.7). Pa(Ci) est calculée pour chaque concept de O et de O' avec les trois hypothèses présentées ci-dessus. Tab. 4.2 montre que les trois mesures S_p, S_s et S_g ne donnent pas la même valeur de similarité pour certains couples de O. En plus, l'ordre des similarités des couples n'est pas le même pour S_P et les deux autres mesures S_s et S_g : $S_P(C7, C4)_O < S_P(C7, C9)_O$ alors que $S_s(C7,C4)_O > S_s(C7,C9)_O$, par exemple.

La modification de O (passage de O à O') a fait changer la valeur des similarités S_s et S_P pour certains couples. En plus, l'ordre stricte des valeurs n'est pas conservé par S_s et S_g : $S_s(C7,C4)_O > S_s(C7, C9)_O$ alors $S_s(C7, C4)_{O'} < S_s(C5, C6)_{O'}$ et $S_g(C4,C9)_O > S_g(C8, C9)_O$ alors $S_g(C4, C9)_{O'} < S_g(C8, C9)_{O'}$, par exemple.

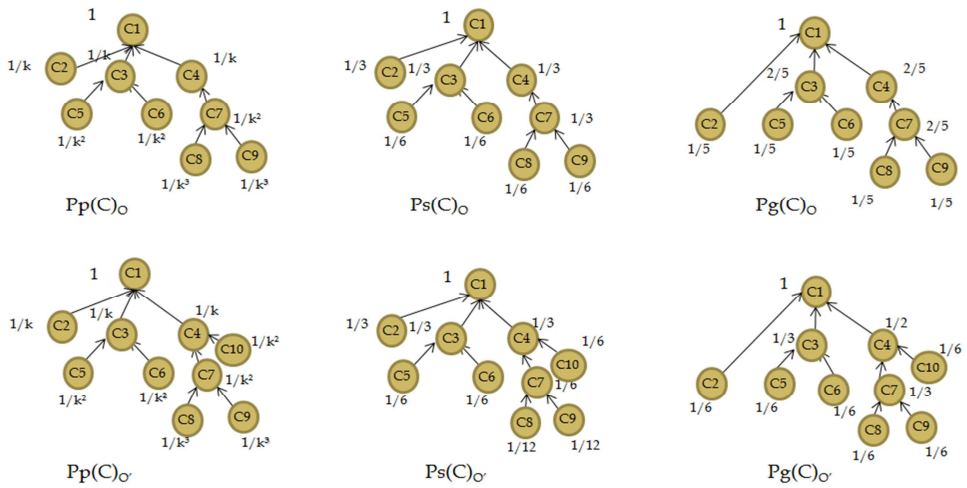

Fig. 4.7 - P_O(Ci) et $P_{O'}$(Ci) pour chaque concept de O et de O'

O	Sp	Ss	Sg
(C4, C9)	0,33	0,61	0,57
(C4, C7)	0,50	1	1
(C7, C9)	0,67	0,61	0,57
(C8, C9)	0,50	0,44	0,40
(C3, C6)	0,50	0,61	0,57
(C5, C6)	0,33	0,44	0,40
O'	Sp	Ss	Sg
(C4, C9)	0,33	0,44	0,39
(C4, C7)	0,50	0,61	0,63
(C7, C9)	0,67	0,72	0,61
(C8, C9)	0,50	0,56	0,44
(C3, C6)	0,50	0,61	0,61
(C5, C6)	0,33	0,44	0,44
(C4, C10)	0,25	0,61	0,39
(C10, C7)	0,33	0,44	0,32
(C10, C8)	0,25	0,35	0,24

Tab. 4.2 -Valeurs des trois mesures pour certains couples de O et O'

Suite à ces constations, nous nous sommes posés trois questions : (1) Pour une application nécessitant une mesure sémantique et utilisant une ontologie donnée, est-ce qu'il y a une mesure qui est la plus adaptée à elle ? ; (2) Si oui, comment peut-on choisir cette mesure ? ; (3) Comment les résultats de cette application varient-ils si l'ontologie utilisée change et comment faut-il réaliser ce changement pour minimiser ces variations ? Cette dernière question considère le cas où un changement dans une ontologie pourrait se réaliser de plusieurs façons. Par exemple, dans l'extrait de Fig. 4.6, le concept « Hydravion » peut être rajouté comme une spécialisation de « Avion », de « Bateau », des deux concepts « Moyen_tansport_maritime » et

« Moyen_de_Transport_aérien » ou de « Moyen_de_transport ». Evidemment, la spécialisation qui apporte le plus de connaissance à cette ontologie est la troisième.

Mais tout d'abord, est-ce vraiment gênant pour une application que des mesures ne donnent pas les mêmes valeurs ou ne donnent pas le même ordre des valeurs ? Est-ce gênant qu'une mesure change de valeur ou d'ordre si l'ontologie sur laquelle est appliquée change ?

L'écart entre les résultats de deux mesures pour une application donnée (1^{er} cas) ou la variation des résultats d'une mesure suite à un changement dans une ontologie ($2^{ème}$ cas) peuvent être évalués avec plusieurs critères. Nous proposons les critères suivants : (1) le nombre de paires de concepts pour lesquelles les valeurs de deux mesures sont différentes (1^{er} cas) ou pour lesquelles la valeur d'une mesure a changé ($2^{ème}$ cas), (2) le nombre de paires de concepts pour lesquelles l'écart entre les valeurs de deux mesures dépassent un certain seuil (1^{er} cas) ou pour lesquelles la valeur d'une mesure a varié de plus d'un certain seuil ($2^{ème}$ cas), et (3) le nombre de deux paires de concepts pour lesquelles les valeurs de deux mesures ont changé d'ordre (1^{er} cas) ou pour lesquelles la valeur d'une mesure a changé d'ordre ($2^{ème}$ cas).

Cet écart sera palpable par une application, si les résultats de cette application changent en utilisant la première ou la deuxième mesure (1^{er} cas) ou suite à un changement dans une ontologie ($2^{ème}$ cas) . La pertinence du critère de l'évaluation de cet écart dépend de l'objectif de cette application. Par exemple, le troisième critère est pertinent pour les applications où la valeur absolue de la mesure n'est pas très importante et où on cherche plutôt l'objet le plus similaire à un objet donné. Ceci implique que dans une recherche d'information pointue ou dans une application précise d'aide à la décision où on cherche par exemple la protéine la plus similaire à une protéine donnée, la différence des ordres des mesures ou la variation de l'ordre d'une mesure peut nuire aux résultats de cette application. Par contre, si on cherche à déterminer un ensemble d'objets le plus proche à un objet donné (par exemple, une recherche sur Google), la différence des ordres des valeurs entre mesures (ou la variation de l'ordre d'une mesure) jusqu'à un certain degré ne nuit pas vraiment aux résultats de l'application.

Dans le cadre du travail de Post Doc de L. Chavin [2012], nous avons commencé une étude portant sur la variation d'une mesure en fonction de l'évolution d'une ontologie. Nous avons cherché à déterminer pour une famille de mesures (caractérisée par son hypothèse d'approximation d'IC) et un type d'ontologie (caractérisé par des propriétés à déterminer), le type de changement à réaliser pour minimiser la variation des résultats de cette mesure. Pour la famille de mesures avec l'hypothèse $H_{yO}1$, P_p et par conséquent IC_p ne changent pas si on rajoute ou on enlève une feuille (un changement élémentaire). Cette famille est robuste pour des changements élémentaires dans une ontologie. D'ailleurs, c'est la mesure la plus souvent utilisée. Cependant, son hypothèse est rarement respectée, comme nous l'avons mentionnée précédemment.

Pour la famille de mesures avec l'hypothèse $H_{yO}2$ (*i.e.* le nombre d'instances d'un concept est divisé par le nombre de ses fils) P_s et par conséquent IC_s ne changent que pour les concepts et leurs subsumants dont le frère a été supprimé ou rajouté. Par conséquent, seulement la similarité de chacun de ces concepts avec les autres concepts de l'ontologie pourrait changer. Elle augmente, si les frères du concept supprimé sont des feuilles. La variation de cette famille de mesures est localisée et bien maitrisée. Mais cette famille est moins robuste que la première.

Pour la famille de mesures avec l'hypothèse $H_{yO}3$ (*i.e.* une distribution uniforme des instances de l'ontologie sur ses feuilles) la détermination de la variation de cette famille n'étant pas, évidente, nous avons réalisé une étude expérimentale exploratoire [Chauvin, 2012]. Nous avons déterminé des critères pour caractériser des ontologies et des changements et nous avons choisi le $3^{ème}$ critère normalisé portant sur le changement d'ordre (noté ChON). Nous avons généré un jeu de données avec des ontologies et des changements sur ces ontologies (40 ontologie/changement)

et nous y avons appliqué la méthode k-means afin de déterminer des classes spécifiques d'ontologies et de changements en fonction de la variation d'une mesure. Cependant, cette étude n'a pas révélé des classes pertinentes. Nous avons continué l'étude en scindant l'ensemble des couples ontologie/changement en deux classes en fonction de la valeur du ChON (nulle ou non). Nous avons déterminé les cas impliquant une variation nulle. En l'occurrence, ces cas concernent (1) les changements qui ont porté sur une feuille qui est un fils unique ; (2) les changements qui ont été effectués sur une feuille liée directement à la racine ; ou (3) les cas où pour deux valeurs successives de la mesure, la somme de l'augmentation maximum de la première valeur et la diminution maximum de la deuxième valeur est toujours inférieure à l'écart initial entre ces deux valeurs. Dans cas, il n'y a pas de changement d'ordre pour tous les couples de l'ontologie. N'ayant pas pu expliquer ce dernier cas, nous avons orienté notre approche vers l'étude de la variation des résultats d'une mesure en fonction de la distribution de ses valeurs et son lien avec la distribution des fils des concepts d'une ontologie.

4.9 Cadre unifiant des mesures sémantiques de comparaison d'objets

Dans les sections précédentes, nous avons proposé une forme unifiée pour la définition des mesures sémantiques pour la comparaison d'objets annotés chacun par un concept, un ensemble de concepts ou un graphe sémantique d'une ontologie. Cette forme unifiée couvre des mesures sémantiques avec des objectifs différents (similarité, inclusion, généralisation, *etc.*). Elle utilise une approximation du contenu informationnel des données qui explicitent la sémantique des artefacts de cette ontologie. Nous avons discuté et proposé différentes approximations du contenu informationnel d'un concept, d'un groupe de concepts et d'un graphe sémantique.

Nous avons regroupé ces résultats dans un cadre paramétrable pour construire un environnement permettant de définir des mesures sémantiques pour un contexte donné (une ontologie et les données caractérisant ses artefacts) et un objectif spécifique. Ce cadre est aussi un moyen pour aider à reformuler certaines mesures existantes, bien comprendre leur sémantique et bien interpréter leurs résultats afin d'aider à en choisir une.

Ce cadre est défini avec une fonction MSCO (Mesure Sémantique de Comparaison d'Objets) et 6 paramètres OC(α ,β), Ai, TD, ICa et h_{ICa} [Harzallah&Berio, 2015]. La fonction MSCO est une mesure sémantique paramétrable de comparaison d'objets, définie comme suit :

$$\mathrm{MSCO}: \mathrm{OB} \times \mathrm{OB} \longrightarrow \Re^{+}$$

$$\mathrm{MSCO}(O1, O2) = \frac{\mathrm{IC}^{\cap}(\mathrm{Ai}(O1), \mathrm{Ai}(O2))}{\alpha \mathrm{IC}^{-}(\mathrm{Ai}(O1), \mathrm{Ai}(O2)) + \beta \mathrm{IC}^{-}(\mathrm{Ai}(O2), \mathrm{Ai}(O1)) + \mathrm{IC}^{\cap}(\mathrm{Ai}(O1), \mathrm{Ai}(O2))}$$

où OB représente l'ensemble des objets à comparer

MSCO est instanciée en choisissant les valeurs des sept paramètres du cadre.

OC(α, β) (**Objectif de la Comparaison).** Une MSCO doit avoir un objectif de comparaison. L'objectif est défini en choisissant les paramètres α et β. Comme nous avons vu précédemment, la famille des mesures avec $\alpha = \beta$ sont des similarités dont l'objectif est d'évaluer une

équivalence ou une égalité entre deux objets. La famille de mesures avec $\alpha = 0$ ou $\beta=0$ sont des mesures d'inclusion.

Type d'annotation **Ai.** MSCO compare des objets selon le type de leur annotation avec une ontologie. Nous distinguons trois types d'annotations Ai d'un objet O : Ac est une annotation par un concept, As est une annotation par un ensemble de concepts, et Ag est une annotation avec un graphe sémantique. Il est évident que si on souhaite prendre en compte la complexité d'un objet, ses différents aspects et sa structure, il faut l'annoter avec un graphe sémantique. Par contre, s'il est caractérisé par différents aspects qui ne sont pas liés entre eux dans ce cas, une comparaison utilisant une annotation par un ensemble de concepts est suffisante. Enfin, l'annotation par un seul concept est souvent utilisée pour comparer des objets simples.

Type des Données (TD). MSCO utilise des données portant sur les concepts d'une ontologie et les liens entre eux. Ces données peuvent être intentionnelles, telles que les propriétés des concepts et des relations taxonomiques ou extensionnelles, telles que les occurrences ou les instances des concepts.

Approximation du contenu informationnel (ICa). Pour quantifier le contenu informationnel commun à deux objets et celui distinctif d'un objet par rapport à l'autre (respectivement $IC^{\cap}$ et IC^{-}), nous avons proposé plusieurs approximations possibles en fonction du type des données disponibles et du type de l'annotation utilisée. Chaque approximation engendre une hypothèse sur les artefacts de l'ontologie et/ou les données utilisées.

Hypothèse de l'approximation de IC (h_{ICa}). A chaque MSCO, une hypothèse est associée pour définir une approximation de $ICa^{\cap}$ et ICa^{-}. Une hypothèse dépend et porte sur les données disponibles (TD) pour approximer le contenu informationnel. Nous avons identifié dans la section 4.4 quelques hypothèses d'approximation.

Nous sommes les premiers à avoir proposé un cadre pour comparer et analyser des mesures sémantiques et en définir si nécessaire une nouvelle. Sanchez&Batet [2011] ont montré que plusieurs mesures sémantiques de similarité peuvent être reformulées en utilisant une réécriture du contenu informationnel. Leurs travaux ressemblent beaucoup à une partie de ce qui a été présentée dans [Blanchard *et al.* 2008].

Cross *et al.* [2013] ont aussi proposé de reformuler les mesures basées sur les propriétés en utilisant le contenu informationnel, dans le cadre des théories des ensembles flous. Nous avons étudié les liens entre les mesures définies en analyse des données et leur analogie avec les mesures sémantiques [Blanchard *et al.* 2008b].

Harispe *et al.* [2014] ont défini récemment un cadre pour généraliser la définition des mesures de similarité sémantique pour deux concepts, utilisant la structure taxonomique. Ils ont proposé une forme générale qui regroupe les deux familles de mesure σ_β[39] et σ_α, définies en analyse des données [Blanchard *et al.* 08b]. Ils ont utilisé six fonctions pour définir une mesure : ρ pour définir la représentation sémantique d'un concept ou d'un ensemble de concepts ; $\Psi(c_1, c_2)$ pour quantifier la partie commune de la représentation sémantique de deux concepts, ce qui est équivalent dans notre cadre à $ICa^{\cap}(c_1, c_2)$; $\Phi(c_1, c_2)$ pour quantifier la représentation sémantique qui différencie un concept d'un autre, ce qui est équivalent dans notre cadre à $ICa^{-}(c_1, c_2)$; $\zeta(c_1, c_2)$ pour quantifier la représentation sémantique d'une ontologie qui ne fait partie ni de c_1 ni de c_2 ; $\theta(c1)$ qui est une fonction pour estimer la spécificité d'un concept dans une ontologie, dans notre cadre ICa permet de définir la spécificité d'un concept ; Θ pour définir le degré de spécificité de la représentation sémantique d'un concept, il a été défini par la somme des spécificités des concepts qui appartiennent à sa représentation, ce qui ne nous semble pas correct

[39] Cette famille de mesure a été discutée dans la section 4.8

(cf. section 4.6.3). Enfin, ils ont utilisé les notions de $\Phi(c_1, c_2)$ et $\Psi(c_1, c_2)$ pour reformuler plusieurs mesures existantes. Par ailleurs, ils proposent de choisir une mesure en faisant varier α, β et z (des coefficients utilisés dans leur forme unifiée des mesures) et d'utiliser leur plateforme pour comparer ses résultats au jugement humain.

De même, Mazandu&Mulder [2013b] et Gaston *et al.* [2016] ont montré que la majorité des mesures peut être reformulée sous la forme d'une fraction où le numérateur est une fonction de l'information commune entre deux concepts et le dénominateur est une fonction linéaire de ce que différencie chaque concept de l'autre concept et de leur partie commune. Pour eux, le dénominateur est un moyen pour normaliser le résultat de la mesure. Pour nous, ce dénominateur fait partie de la définition d'une mesure de comparaison.

4.10 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux mesures sémantiques comme un moyen pour gérer l'hétérogénéité des données et connaissances. Plusieurs mesures sémantiques sont proposées dans la littérature. Elles ont des formes différentes et utilisent souvent la structure d'une taxonomie seule ou avec un corpus de textes. Elles intègrent parfois une approximation du contenu informationnel dont le lien avec la définition originelle du contenu informationnel n'est toutefois pas explicité. Leur précision est souvent évaluée à l'aide d'un jeu de données en comparant la corrélation de leurs résultats avec le jugement humain. Cependant, cette précision peut changer si on modifie le contexte de leur utilisation. L'analyse des résultats des mesures en fonction de leur contexte ou en fonction de leurs caractéristiques intrinsèques n'a pas été vraiment étudiée dans la littérature.

Nous avons proposé un cadre prenant en compte différents paramètres dont dépend une mesure sémantique et permettant de définir une nouvelle mesure ou reformuler une mesure existante. Nous avons particulièrement mis en évidence l'existence d'une hypothèse inhérente à chaque mesure sémantique portant sur les artefacts de l'ontologie utilisée, dont la validité détermine la précision de cette mesure. Dans notre cadre, trois familles principales de mesures sémantiques sont à distinguer selon le type d'annotation des objets à comparer : annotation par un concept, annotation par un ensemble de concepts ou annotation par un graphe sémantique. Les trois familles sont définies selon la même approche, basée sur une approximation du contenu informationnel commun à deux objets et sur celle du contenu informationnel distinctif d'un objet par rapport à l'autre.

La famille de mesures sémantiques utilisant une annotation avec un graphe sémantique, d'après nos connaissances, est innovante et pertinente. Elle est innovante parce qu'il n'y a pas encore de mesures qui comparent deux graphes sémantiques à l'aide de leur graphe sémantique commun défini avec une ontologie. Elle est pertinente parce qu'elle prend en compte en plus de la sémantique des aspects d'un objet, la sémantique des liens entre eux.

Nous avons appliqué notre cadre pour la définition de mesures sémantiques dans plusieurs domaines : (1) dans le domaine de l'ingénierie des modèles pour comparer deux constructs d'un langage où ces derniers ont été annotés par des groupes de concepts et ensuite par des graphes sémantiques; (2) dans le domaine des métiers d'entreprise pour l'interopérabilité des systèmes hétérogènes où les données d'un système ont été annotées par des concepts uniques ; et (3) sur des réseaux sociaux de média pour comparer des profils d'utilisateurs annotés par un ou plusieurs concepts afin de détecter des communautés.

En ce qui concerne nos perspectives de recherche, nous allons tout d'abord travailler sur trois axes d'amélioration de nos travaux. Nous envisageons de : (1) intégrer la prise en compte du

phénomène central dans les familles de mesures de comparaison utilisant une annotation par un ensemble de concept ou par un graphe sémantique, (2) approfondir notre étude de la variation des résultats d'un type de mesure en fonction du changement réalisé dans un type d'ontologie donnée, et (3) déterminer une approche pour le choix d'une mesure, qui s'adapte le mieux à une ontologie donnée. L'objectif des deux derniers axes d'amélioration nous semble très pertinent pour le choix d'une mesure sémantique la plus adaptée à une application et une ontologie dans le web des données, où le nombre d'applications de comparaison croit et les ontologies utilisées évoluent.

Notre dernière perspective est la définition d'une nouvelle famille de mesures sémantiques, en utilisant notre cadre, pour la détection de communautés dont les centres d'intérêt de leurs individus sont annotés par une ontologie. En effet, nous souhaitons adapter la famille de mesures sémantiques de comparaison de deux ensembles de concepts à la comparaison de deux communautés dans un réseau social en considérant le nombre d'individus ayant le même centre d'intérêt dans une même communauté [Ben Amor *et al.* 2016].

Chapitre 5 : Conclusion, Projet de Recherche et Perspectives

Nos travaux de recherche réalisés durant 16 années au LS2N (LINA, jusqu'à décembre 2016) ont porté principalement sur le développement des modèles sémantiques, des cadres ou des méthodes dont certaines utilisent des techniques de raisonnement logique ou de traitement automatique des langues. Plus précisément, nos contributions s'articulent autour de 3 axes : (1) l'ingénierie des compétences des ressources humaines, (2) la construction et la validation semi-automatiques d'ontologie à partir des textes, et (3) les mesures sémantiques de comparaison d'objets. Dans ces trois axes, nous nous sommes intéressés aux ontologies de taille moyenne et offline. Toutefois, nos travaux pourraient participer à répondre au challenge des masses de données et à la construction d'ontologie de grande taille et à son exploitation.

Tout d'abord, notre architecture intégrante pour l'ingénierie des compétences qui permet d'inventorier et de structurer des ressources des connaissances et des techniques d'ingénierie des connaissances peut s'étendre à une architecture intégrante pour l'ingénierie des connaissances à partir de tout type de données. Dans cette architecture, les processus d'ingénierie des connaissances correspondront aux processus de recherche d'information, de recommandation, de prédiction… en plus de ceux portant sur l'extraction, la modélisation, l'exploitation et le maintien des connaissances. D'une façon générale, ces processus participeront à l'automatisation des différents processus d'une organisation [André *et al.* 2017]. Des techniques de fouille de masses de données de différents types seront à intégrer dans l'ontologie de l'ingénierie des connaissances de cette architecture.

Notre approche pour la construction et la validation d'ontologie pour l'annotation d'objets constitue une contribution pertinente pour le développement d'une approche de construction d'ontologie pour l'annotation des objets du web. En effet, l'utilisation d'une représentation structurelle/fonctionnelle d'un objet pour son annotation aide à améliorer le résultat de ce processus d'annotation sur le web. En plus, l'enrichissement d'une ontologie selon les besoins en annotation d'objets limite la taille de cette ontologie tout en s'adaptant à l'évolution du web.

Notre approche d'optimisation de l'intervention humaine pour la validation d'ontologie ou les dépendances de validation que nous avons proposées pour l'identification des problèmes dans une ontologie sont des contributions pertinentes pour mieux gérer la complexité de la validation d'ontologie de grande taille.

Enfin, notre cadre unifiant des mesures sémantiques de comparaison d'objets et nos réflexions en cours sur la variation d'une mesure sémantique ont leur place pour la définition ou le choix de mesures sémantiques pour les applications du web des données.

Les quatre premiers chapitres de ce manuscrit ont présenté une synthèse de nos activités de recherche passées et des résultats obtenus. Nous présentons maintenant les perspectives de nos travaux de recherche que nous souhaitons traiter en priorité. Bien sûr, ces perspectives s'inscrivent pleinement dans les objectifs de l'équipe DUKe du LS2N.

Projet de Recherche et Perspectives

Actuellement, pour relever le challenge de masse de données et ses applications, plusieurs thèmes de recherche sont traités aux niveaux national ou international, chacun seul ou d'une façon interreliée. Nous pouvons en citer certains qui sont liés à nos perspectives :

1. Combinaison des méthodes du TAL, de fouille des données, d'apprentissage automatique et de raisonnement logique pour la construction d'ontologie [LeCun, 2016], [Lefever, 2016];
2. Enrichissement/Peuplement d'ontologie ou de ressources sémantiques en utilisant Wikipédia et DBpédia, [Lopez *et al*. 2014], [Nebhi, 2013], [Booshehri&Luksch, 2015], [Kliegr, 2015], [Haidar-Ahmad *et al.* 2016];
3. Amélioration de l'annotation des documents du web avec l'ontologie DBpédia ou d'autres ontologies [Suchanek, 2014], [Alec, 2016];
4. Développement d'outils pour la fouille de textes, pouvant être adaptés à la construction et au peuplement d'ontologies ou de ressources sémantiques. Par exemple, nous pouvons citer des packages développés pour R pour la fouille des textes [Uslu *et al.* 2017].

Dans notre projet de recherche, nous continuerons à travailler sur la construction et la validation d'ontologie et son exploitation, tout en participant à relever le challenge des masses de données. Nous envisageons d'intégrer dans nos travaux futurs des techniques de traitement automatique de langues, de fouille des données, d'apprentissage automatique, et de raisonnement logique. Nous envisageons également d'augmenter la quantité de textes à partir de laquelle la construction sera réalisée, en allant progressivement vers une masse importante de textes à partir du web, pour aboutir à une démarche de construction, d'enrichissement et de validation d'ontologie à la volée, à partir du web et pour le web (à court –moyen terme).

Par ailleurs, nous envisageons de continuer à travailler sur les mesures sémantiques de comparaison en appliquant les résultats de nos travaux pour la comparaison de nouveaux types d'objets, voire à des objets du web et avec des ontologies du web (quand l'occasion se présente).

Mais tout d'abord, nous envisageons de valider certains de nos résultats sur des jeux de données conséquents, voire réels (à court terme). Plus particulièrement, il s'agit (1) de la validation des dépendances entre problèmes et des anti-patrons partiels pour l'aide à l'identification du problème de « Contradiction sociale » et de leur intégration dans notre démarche globale de validation d'ontologie et (2) de la validation de notre mesure sémantique de comparaison d'objets annotés par des graphes sémantiques. En plus, nous envisageons de valoriser et maintenir les prototypes développés (*e.g.* le prototype pour la comparaison d'objets annotés par un concept ou un ensemble de concepts, le prototype pour la comparaison d'objets annotés par un graphe sémantique, Text2Onto pour le français).

Nous exposons dans la suite quatre perspectives, chacune pouvant correspondre (ou correspond déjà) à un ou plusieurs sujets de thèse. Elles sont présentées selon leur ordre d'importance dans notre projet de recherche.

Perspective 1. *Développement d'une approche semi-automatique de construction et validation intégrées d'ontologie à partir d'un nombre important de textes, utilisant une ontologie noyau formelle (travaux en cours ou à court et à moyen terme).*

Nous avons considéré dans le projet KIFANLO et ISTA3 le processus de conceptualisation semi-automatique d'ontologie à partir de textes. Nous avons remarqué le manque de méthode et outil efficaces pour réaliser ce processus. Dans cette première perspective, nous envisageons de développer une approche semi-automatique de construction incrémentale basée sur les principes suivants :

- Construction d'une ontologie modulaire structurée selon une ontologie noyau formelle : chaque concept noyau représente la racine d'un module de cette ontologie ;
- Construction semi-automatique incrémentale : tous les textes ne sont pas traités en une seule fois. Ils sont plutôt classés et traités par ordre croissant du degré de spécialisation de leur

contenu [Benard *et al.* 2014], [Harzallah *et al.* 2014]. A chaque itération, l'idée est de rajouter dans l'ontologie à développer des concepts de degré de spécialisation plus important que celui des concepts déjà présents dedans. Néanmoins, ce principe requiert une classification préalable des textes par degré de généricité de leur contenu, ce qui n'est pas une tâche simple ;

- Validation semi-automatique basée sur une ontologie noyau formelle. Ce principe utilise les axiomes d'une ontologie noyau formelle et les règles inférées à partir de celle-ci pour valider les concepts ou les relations ajoutées (cf. section 3.3.1.3 et section 3.3.2). Ce type de validation est réalisée après chaque itération en utilisant les artefacts validés dans l'itération précédente ;
- Validation de l'ontologie par l'utilisateur après chaque itération. Ce principe permet de gérer la complexité de la validation et d'améliorer la qualité de l'enrichissement.

Dans cette perspective, nous envisageons développer une approche de construction d'ontologie qui combinera des techniques de TAL, de fouille des données, d'apprentissage automatique et de raisonnement (thèmes de recherche n°1 et n°4, cités ci-dessous).

Dans le cadre de cette perspective, nous co-encadrons la thèse de A. Alaa Eddine depuis décembre 2016 avec Giuseppe Berio et Nicolas Bechet (IRISA, Vannes) et Ahmad Faour (Université du Liban). Cette thèse porte principalement sur l'apprentissage de patrons pour l'extraction de relations à partir des connaissances profondes (i.e. arbres de dépendances) et d'une ontologie noyau formelle.

En outre, dans le cadre de la thèse Z. Xu, en co-encadrement avec Fabrice Guillet (membre de l'équipe Duke), nous envisageons de développer des nouveaux modèles de fouille de motifs enrichis en combinant des modèles probabilistes de "topic modeling" (en l'occurrence l'approche LDA), des règles d'association et des motifs sémantiques, afin d'enrichir l'ontologie DBPédia et l'annotation des documents de Wikipédia, et améliorer ainsi la recommandation de ces documents. Pour évaluer nos résultats, nous envisageons d'appliquer nos modèles à la construction d'une ontologie à partir d'un corpus de Wikipédia portant sur un domaine particulier. Ensuite, pour ce domaine, nous comparerons nos résultats à DBpedia et à l'annotation existante par DBpedia des documents de Wikipédia (axes de recherche n°2 et n°3, cités ci-dessous).

Nous encadrons également le travail de master de S. Hedhli pour la construction semi-automatique d'ontologie combinant des techniques de classifications non supervisées ou supervisées et l'utilisation d'une ontologie noyau, à partir d'un corpus de textes.

Nous co-encadrons le travail de master de A. Méchergui pour le développement d'une approche de construction semi-automatique d'ontologie combinant la méthode LDA et l'utilisation d'une ontologie noyau, à partir d'un corpus de textes.

***Perspective 2**. Développement d'une approche semi-automatique de validation d'ontologie intégrée à la construction et dirigée par les problèmes pouvant nuire à la qualité d'ontologie (en cours et à court terme).*

Nous avons identifié dans nos travaux une liste de dépendances de validation entre les problèmes d'une ontologie. Ces dépendances permettront d'optimiser l'intervention humaine dans un processus de validation d'ontologie. D'un autre côté, nous avons défini des anti-patrons partiels pour l'aide à l'identification du problème de « Contradiction sociale ».

Nous envisageons tout d'abord d'identifier des types de dépendances de validation et de les formaliser. Ensuite, nous envisageons d'étendre notre liste de dépendances à des nouvelles dépendances. Puis, nous allons définir des règles pour déterminer un ordre de vérification et de correction des problèmes optimisant le processus de validation, en minimisant l'intervention de

l'acteur social. L'ordre défini sera modélisé sous la forme d'une machine à état qui sera adaptée à une ontologie en fonction des problèmes qui pourraient nuire à sa qualité.

Enfin, nous envisageons d'intégrer l'ordre proposé pour l'identification et la correction des problèmes et les anti-patrons partiels développés (des anti-patrons partiels pour d'autres problèmes de type « social » sont à développer) dans une approche de construction et de validation intégrées d'ontologie.

Pour cette perspective, nous envisageons de chercher un benchmark pour valider l'ordre proposé pour l'identification et la correction des problèmes. L(es) ontologie(s) de ce benchmark doiv(en)t comprendre des axiomes autres que les axiomes de subsumption. Le benchmark doit comprendre aussi les corpus à partir desquels ces ontologies ont été construites.

Perspective 3. *Amélioration et validation de notre cadre unifiant des mesures de comparaison sémantiques d'objets et son application à la comparaison des objets du web (à Court et à Moyen terme).*

Tout d'abord, comme nous avons signalé dans le chapitre 4, nous souhaitons améliorer notre cadre unifiant en développant une approche pour le choix d'une mesure en fonction de son hypothèse qui s'adapte au mieux à une ontologie donnée. Cette approche est très pertinente pour le choix d'une mesure sémantique pour une application et une ontologie dans le web, où le nombre d'applications de comparaison et les ontologies utilisées croissent et évoluent rapidement.

Ensuite, nous souhaitons adapter la famille de mesures sémantiques de comparaison de deux ensembles de concepts à la comparaison de deux communautés dans un réseau social en considérant le nombre d'individus ayant le même centre d'intérêt dans une même communauté [Ben Amor *et al.* 2016].

A moyen terme, nous envisageons l'application de notre cadre unifiant à la comparaison des objets du web. Ceci requiert, entre autres, le choix d'une ontologie d'annotation d'objets, la prise en compte de son évolution, le choix d'une mesure sémantique, et la prise en compte du processus d'annotation sur le web et l'évolution de ces objets (thème de recherche n°3, cité ci-dessous).

Perspective Enseignement-Recherche : *Exploitation des ontologies pour l'interopérabilité des ERPs et des applications informatiques du hall de production du département QLIO de l'IUT de Nantes (à moyen terme)*

Dans le département QLIO, plusieurs logiciels ERPs sont utilisés dans des travaux pratiques en gestion de production, ordonnancement, gestion de coût, *etc.* D'autres logiciels sont utilisés dans le hall de production pour la planification de la production, le lancement et le suivi. Il nous semble intéressant, dans un souci de montrer à l'étudiant un processus complet informatisé de gestion et de pilotage de la production, de faire interopérer ces différents logiciels, à l'aide d'une ontologie portant sur le domaine de pilotage de la production. Nous souhaitons donc développer ce type d'ontologie à partir des « Supports d'Aide » des logiciels et participer au développement d'un vrai système d'interopérabilité de ces différents logiciels.

Nous avons déjà traité certaines données de l'ERP CEGID Business pour construire une ontologie dans la thèse de T. Gherasim. Nous avons également co-encadré avec Nasser Mébarki, enseignant-chercheur du département QLIO Nantes, un projet tutoré en licence LGPI sur l'interopérabilité de deux logiciels de GPAO (i.e. Prélude et Just In Time).

Annexe. Liste des problèmes analysés et classés dans notre typologie

Problèmes traités dans les travaux de Poveda *et al.*

- Creating Polysemious elements
- Creating synonyms as classes
- Creating the relationship "is" instead of using "rdfs :subClassOf", "rdfs :type" or "owl :sameAs"
- Creating unconnected ontology elements
- Wrong inverse relationship
- Including cycles in the hierarchy
- Merging different concepts in the same class
- Missing annotations
- Missing basic information
- Missing disjointness
- Missing domain or range in properties
- Missing equivalent property
- Missing inverse relationships
- Misusing "owl :AllValuesFrom"
- Misusing "not some" and "some not"
- Misusing primitive and defined classes
- Specializing a hierarchy exceedingly
- Specifying the domain or range exceedingly
- Swapping intersection and union
- Misusing ontology annotations
- Using a miscellaneous class
- Using different naming criteria in the onotlogy
- Using incorrectly ontology elements
- Using recursive definition
- Define a relationship inverse to itself
- Defining inverse relationships for a symmetric one
- Defining wrong equivalent relationship
- Defining wrong symmetric relationship
- Definind wrong transitive relationship
- Missing equivalent classes
- Defining wrong equivalent classes
- Several classes with the same label
- Creating a property chain with just one property

Problèmes traités dans les travaux de Roussey *et al.*

- Antipattern AndIsOr
- Antipattern OnlynessIsLonelyness
- Antipattern UniversalExistence
- Antipattern EquivalenceIsDifference

- Antipattern SynonymOfEquivalence
- Antipattern SumOfSome
- Antipattern SomeMeansAtLeastOne
- Guideline DisjointnessOfComplement
- Guideline Domain&CardinalityConstraints
- Guideline GroupAxioms Guideline MinIsZero

Problèmes traités dans les travaux de Bühmann *et al.*

- Partition error
- Semantic inconsistency
- Having both a class and its complement as super condition
- Having a super condition that is assume to be disjoint with owl :Thing
- Having a super condition that is an existential restriction that has a filler which is disjoint with the range of the restricted property
- Having an universal restriction with owl :Nothing as the filler and a must existing restriction along property relationships
- Having more than allowed existential restrictions
- Having super conditions containing conflicting cardinality restrictions
- Inconsistence from other sources

Problèmes traités dans les travaux de Fahad *et al.*

- Lazy concept
- Chain of inheritance
- Lonely disjoint
- Property clump

Problèmes traités dans les travaux de Qazi&Abdul

- Redundancy of Subclass
- Redundancy of Disjoint Relation

Références

[Abiteboul *et al.* 2017] Abiteboul S., M. Arenas, P. Barceló, M. Bienvenu, D. Calvanese, C. David, R. Hull, E. Hüllermeier, B. Kimelfeld, L. Libkin, W. Martens, T. Milo, F. Murlak, F. Neven, M. Ortiz, T. Schwentick, J. Stoyanovich, J. Su, D. Suciu, V. Vianu, K. Yi. Research Directions for Principles of Data Management. Dagstuhl Perspectives Workshop. Dagstuhl Publishing, Germany, 2017.

[Abiteboul, 2012] Abiteboul S. Sciences des données: de la logique du premier ordre à la Toile. Leçon inaugurale au Collège de France prononcée le jeudi 8 mars, 2012 (http://books.openedition.org/cdf/529).

[Aimé&Charlet, 2014] Aimé X, J Charlet. Knowledge Engineering or Conformism Engineering? In: Bergenti F, Cabri G, editors. Enabling Technologies: Infrastructure for Collaborative Enterprises. 23rd International WETICE IEEE pp.399–404, 2014.

[Amardeilh&Damljanovic, 2009] Amardeilh F., D. Damljanovic. Du texte à la connaissance : annotation sémantique et peuplement d'ontologie appliqués à des artefacts logiciels. 20èmes journées francophones d'ingénierie des Connaissances. pp.29, 2009.

[Ambauen *et al.* 2003] Ambauen R., S. Fischer, H. Bunke Graph Edit Distance with Node Splitting and Merging, and Its Application to Diatom Identification. In *IAPR-TC15Workshop on Graph-based Representation in Pattern Recognition*, pp. 95-106, 2003.

[Anaya *et al.* 2008] Anaya V., G. Berio, M. Harzallah, P. Heymans, R. Matulevicius, A.L. Opdahl, H. Panetto, M.J. Verdecho. The Unified Enterprise Modelling Language – Overview and Further Work. Keynote paper in Proceedings of IFAC World Congress, V.17, Part.1, Seoul, Corée, 2008.

[Anaya *et al.* 2010] Anaya V., G. Berio, M. Harzallah, P. Heymans, R. Matulevicius, A. L. Opdahl, H. Panetto, M. J. Verdecho. The Unified Enterprise Modelling Language - Overview and Further Work. Computers in Industry 61(2) pp 99-111, 2010.

[Agrawal *et al.* 1993] Agrawal R., T. Imielinski, A. N Swami. Mining Association Rules between Sets of Items in Large Databases, Proceedings of the 1993 ACM International Conference on Management of Data, Washington, D.C., Peter Buneman and Sushil Jajodia, pp. 207-216, 1993.

[Aries *et al.* 2008] Aries S., B. Le Blanc, JL. Ermine. MASK : une méthode d'ingénierie des connaissances pour l'analyse et la structuration des connaissances. Management et ingénierie des connaissances, modèles et méthodes, Hermes sciences, Traité IC2, Série Management et Gestion des STIC.pp.208, 2008.

[Asgari *et al.* 2015] Asgari R, M. G. Moghadam, M. Mahdavi, A. Erfanian. An ontology-based approach for integrating heterogeneous databases. Open Computer Science, V. 5, I. 1, 2015.

[Aussenac-Gilles *et al.* 2008] Aussenac-Gilles N, S. Després, S. Szulman. The Terminae Method and Platform for Ontology Engineering from Texts. Chapitre de livre Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text, IOS press, pp.192-223, 2008.

[Auer *et al.* 2007] Auer S, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the International Semantic Web Conference (ISWC), volume 4825 of Lecture Notes in Computer Science, pp. 722–735, Springer2007.

[Aussenac-Gilles *et al.* 2014] Aussenac-Gilles N., Charlet J., Reynaud C. Ingénierie des connaissances, Chapitre 20 in Panorama de l'intelligence artificielle. Représentation des

connaissances et formalisation des raisonnements, Volume 1. Eds: Marquis P., Papini O., Prades H. Toulouse : Cepadues Edi., 2014.

[Aussenac-Gilles&Gandon, 2013] Aussenac-Gilles N., Gandon F., From the knowledge acquisition bottleneck to the knowledge acquisition overflow: A brief French history of knowledge acquisition, in Int. J. Human-Computer Stud.es, 71 157–165, 2013.

[Aussenac-Gilles, 2005] Méthodes ascendantes pour l'ingénierie des connaissances Informatique. Mémoire de HDR, Université Paul Sabatier - Toulouse III, 2005.

[Booshehri , &Luksch, 2015] Booshehri M., P. Luksch, An Ontology Enrichment Approach by Using DBpedia, Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics, July 13-15, 2015, Larnaca, Cyprus

[Bachimont et al. 2002] Bachimont, B., A. Isaac, R. Troncy. Semantic commitment for designing ontologies: A proposal. In Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, V 2473 of Lecture Notes in Computer Science, pp 114–121. Springer Berlin Heidelberg, 2002.

[Banko *et al.* 2007] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open In-formation Extraction from the Web. In In the Proceedings of the 20th International Joint Conference on Artificial Intelligence, pp 2670–2676, 2007.

[Baumeister&Seipel, 2005] Baumeister, J., D. Seipel. Smelly owls-design anomalies in ontologies. pp 215–220 of: Proc. of 18th int. florida artificial intelligence research society conf., 2005.

[Baumeister&Seipel, 2010] Baumeister, J. D.Seipel. Anomalies in ontologies with rules. Web semantics: Science, services and agents on the world wide web, 8(1), pp 55–68, 2010.

[Baziz *et al.* 2004] Baziz M. Boughanem M., N. Aussenac-Gilles. The Use of Ontology for Semantic Representation of Documents. In Semantic Web and Information Retrieval Workshop at SIGIR 2004 (SWIR 2004), Sheffield UK, Y. Ding, K. van Rijsbergen, I. Ounis, J. Jose (Eds.), ACM, pp. 38-45, 2004.

[Béchet *et al.* 2012] Béchet N., Cellier P., Charnois T., Crémilleux B. Fouille de motifs séquentiels pour la découverte de relations entre gènes et maladies rares. In IC 2012, 23èmes journées francophones d'Ingénierie des Connaissances, 2012.

[Ben Amor *et al.* 2016] Ben Amor S., L. Ben Romdhane, M. Harzallah. SemMEP : Nouvelle approche sémantique pour la détection des communautés dans un réseau social Actes des 27ème journées francophones d'ingénierie des Connaissances (IC'2016), 2016.

[Benard&Vourch, 2014] Benard K., T. Vourch. Construction semi-automatique d'une ontologie de la pêche. Rapport de projet Transversal avec une entreprise, PolytechNantes, Département Informatique, 2014.

[Benard *et al.* 2014] Benard K., T. Vourch, M. Harzallah I. Tillier Construction semi-automatique et incrémentale d'une ontologie de la pêche à partir des textes. Atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement (IN-OVIVE) de IC'14, 2014 (Article accepté mais non présenté).

[Bendaoud *et al.* 2005] Bendaoud R, Y. Toussaint, A. Napoli. Hiérarchisation des règles d'association en fouille de textes. Revue des Sciences et Technologies de l'Information - Serie ISI : Ingénierie des Systèmes d'Information, Lavoisier, 2005.

[Ben Ishak *et al.* 2011] Ben Ishak, M., P. Leray, N. Ben Amor. A two-way approach for probabilistic graphical models structure learning and ontology enrichment. In Proceedings of the 3rd International Conference on Knowledge Engineering and Ontology Development

(KEOD 2011) part of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management IC3K, pp 189–194, 2011.

[Ben Mustapha et al. 2013] Ben Mustapha N., Aufaure M-A., Baazaoui-Zghal H., Ben-Ghézala H. Query-driven approach of contextual ontology module learning using web snippets, special issue on Database Management and Information Retrieval, Journal of Intelligent Information Systems JIIS, 2013

[Ben Mustapha et al. 2009] Ben Mustapha N., H. Baazaoui Zghal, M.-A. Aufaure, H. Ben Ghezala. Ontology learning from Web: survey and framework based on semantic search, Second International Conference on Web and Information Technologies (ICWIT'09),2009.

[Benelallam *et al.* 2015] Benelallam A., A. Gomez, M. Tisi, J. Cabot. Distributed Model-to-Model Transformation with ATL on MapReduce,Proceedings of 2015 ACM SIGPLAN International Conference on Software Language Engineering, 2015.

[Berio&Harzallah, 2005a] Berio G., M. Harzallah. Knowledge Management for Competence Management, on the online Journal on Universal Knowledge Management (J.UKM), V0, I 1, pp 21-28, 2005.

[Berio&Harzallah, 2005b] Berio G., M. Harzallah, Knowledge Management for Competence Management. Proceeding of the 5th International Conference on Knowledge Management (I-Know05), 2005.

[Berio&Harzallah, 2005c] Berio G., M. Harzallah. De l'ingénierie des connaissances à la gestion des compétences, 16èmes journées francophones d'Ingénierie des connaissances, IC'05, 2005.

[Berio *et al.* 2007] Berio G., M. Harzallah, G.M. Sacco. Portals for Integrated Competence Management. Encyclopaedia of Portal Technology and Applications. Idea Group Inc. 2007.

[Berio&Harzallah, 2007] Berio G., M. Harzallah. Towards an Integrating architecture for competence management, In Computers in Industry, Elsevier, V58, I2, pp. 199-209, 2007.

[Berio *et al.* 2008] Berio G., M. Harzallah, G.M. Sacco. e-HRM – An Integrating Architecture for Competence Management, Encyclopedia of HRIS: Challenges in e-HRM, Idea Group Inc, 2008.

[Berio *et al.* 2011] G. Berio, A. Di Leva, M. Harzallah, G.M. Sacco. Competence management over social networks through dynamic taxonomies. In Encyclopedia of KM2.0. 2011.

[Bézivin, 2006] Bézivin J. *Model Driven Engineering: An Emerging Technical Space* - Generative and transformational techniques in software engineering, Springer Berlin Heidelberg pp 36-64, 2006.

[Bittner&Smith, 2004] Bittner T, B. Smith. Normalizing Medical Ontologies using Basic Formal Ontology. In: Proc. GMDS Innsbruck. pp. 199–201, 2004.

[Blanchard *et al.* 2004] Blanchard J., F. Guillet, R. Gras, H. Briand. Mesurer la qualité des règles et de leurs contraposées avec le taux informationnel TIC. *Revue des Nouvelles Technologies de l'Information.* pp. 287–298, Actes des journées Extraction et Gestion des Connaissances, 2014.

[Blanchard&Harzallah, 2004] Blanchard E., M. Harzallah. Reasoning on competence management, Workshop on Knowledge Management and Organizational Memories of the 16th European conference on Artificial Intelligence (ECAI'04), Valence, pp 22-27, 2004.

[Blanchard *et al.* 2005] Blanchard E., M. Harzallah & H. Briand. Raisonnement en gestion des compétences, Revue des Nouvelles Technologies de l'Information, V3, N 2, pp587-592, 2005.

[Blanchard *et al.* 2005a] Blanchard E., M. Harzallah, H. Briand, P. Kuntz. A typology of ontology-based semantic measures. In the workshop « Enterprise modelling and ontology » (EMO) of the 17the International Conference on Advanced Information Systems Engineering (CAISE'05), Springer Verlag, Riga, 2005.

[Blanchard *et al.* 2006a] E. Blanchard, P. Kuntz, M. Harzallah, H. Briand, A tree-based similarity for evaluating concept proximities in an ontology, in Proc. 10th Conf. Int. Federation Classification Soc., pp 3–11. Springer, 2006.

[Blanchard *et al.* 2006b] Blanchard E., M. Harzallah, P. Kuntz, H. Briand. Une nouvelle mesure sémantique pour le calcul de la similarité entre deux concepts d'une même ontologie, Revue Nationale des Nouvelles Technologies de l'Information, E6, pp. 193-198, Cépaduès Edition, 2006.

[Blanchard *et al.* 2007] Blanchard E., M. Harzallah, P. Kuntz. Vers une classification des similarités basées sur le contenu informationnel des concepts d'une hiérarchie de subsumption, Actes des 18ème journées francophones d'ingénierie des Connaissances (IC'2007), pp.145-156, 2007.

[Blanchard, 2008] Blanchard E. Exploitation d'une hiérarchie de subsomption par le biais de mesures sémantiques. Thèse en informatique de l'université de Nantes, 2008.

[Blanchard *et al.* 2008a] Blanchard E., M. Harzallah, P. Kuntz, H. Briand. Sur l'évaluation de la quantité d'information d'un concept dans une taxonomie et la proposition de nouvelles mesures, Journées francophones d'Extraction et de Gestion des Connaissances (EGC'08), Revue des nouvelles technologies de l'information, pp.127-145, RNTI-E12, Cépa. Ed., 2008.

[Blanchard *et al.* 2008b] Blanchard E., M. Harzallah, P. Kuntz. A generic framework for comparing semantic similarities on a subsumption hierarchy, in Proceedings of the 18th European Conference on Artificial Intelligence (ECAI'2008), IOS Press, pp 20-24, 2008.

[Blomqvist, 2009] Blomqvist E. Semi-automatic Ontology Construction based on Patterns. PhD thesis, Linköping University, Department of Computer and Information Science at the Institute of Technology, 2009.

[Boeres et al. 2004] Boeres M., C. Ribeiro, I. Bloch. A randomized heuristic for scene recognition by graph matching. In WEA 2004, pp. 100–113, 2004.

[Bordea&Buitelaar, 2010] Bordea G., Buitelaar P. Expertise mining. In Proceedings of the 21st National Conference on Artificial Intelligence and Cognitive Science, 2010.

[Bordea et al. 2013] Bordea, G., T. Polajnar, P. Buitelaar. Domain-Independent Term Extraction Through Domain Modelling. In 10th International Conference on Terminology and Artificial Intelligence, 2013.

[Bourigault *et al.* 2004] Bourigault D. N. Aussenac-Gilles, J. Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. Revue d'Intelligence Artificielle. Volume 18, n° 1, pp. 87-110, 2004.

[Bourse *et al.* 2002] Bourse M., M. Harzallah, M. Leclère, F. Trichet. CommOnCV: modeling the competencies underlying a Curriculum Vitae. Proceedings of the 14th International Conference on Software Engineering and Knowledge Engineering (SEKE'2002), ACM Press pp 65-73, 2002.

[Budanitsky&Hirst, 2001] A. Budanitsky, G. Hirst. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures, in *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics*, 2001.

[Buggenhout&Ceusters, 2005] Buggenhout CV, Ceusters W. A novel view on information content of concepts in a large ontology and a view on the structure and the quality of the ontology.Int J Med Inform;74, pp 125–32, 2005.

[Buhmann *et al.* 2011] Buhmann, L., S. Danielczyk, J. Lehmann. D3.4.1 report on relevant automatically detectable modelling errors and problems. Tech. rept. LOD2 - Creating Knowledge out of Interlinked Data, 2011.

[Buitelaar *et al.* 2004] Buitelaar, P., D. Olejnik, M. Sintek. A protege plug-in for ontology extraction from text based on linguistic analysis. dans Proceedings of the 1st European Semantic Web Symposium (ESWS). pp 31–44, 2004.

[Buitelaar *et al.* 2005] Buitelaar P., P. Cimiano, B. Magnini. Ontology Learning from Text : Methods, Applications and Evaluation. Chap. in Ontology learning from text : an overview, IOS Press. pp 3–12, 2005.

[Buitelaar&Eigner, 2008] Buitelaar P., T. Eigner. Topic extraction from scientific literature for competency management. In Personal Identification and Collaborations : Knowledge Mediation and Extraction (PICKME2008), 2008.

[Bunescu&Mooney, 2006] Bunescu R., R. Mooney. Subsequence kernels for relation extraction. In:Weiss, Y., B. Scholkopf, J. Platt, (eds.) Advances in Neural Information Processing Systems MIT Press, 18, pp171-178, 2006.

[Bunge, 1977] Bunge M. Treatise on Basic Philosophy. V3, Ontology I : The furniture of the World. Boston Reidel, 1977.

[Burita *et al.* 2012] Burita L., P.Gardavsky, T. Vejlupek. K-GATE Ontology Driven Knowledge Based System for Decision Support. In Journal of Systems Integration 3(1), pp 19 – 31, 2012.

[Bunke, 1997] Bunke H. On a relation between graph edit distance and maximum common sub-graph. Pattern Recognition Letters 18(8), pp 689-694,1997.

[Bunke&Shearer, 1998] Bunke H., K. Shearer. A graph distance metric based on the maximal common subgraph. In: Pattern Recognition Letters, V. 19 No. 3-4 pp. 255-259, 1998.

[Cameron *et al.* 2010] Cameron D., B. Aleman-Meza, I. B. Arpinar, S. L. Decker, A. P.Sheth. A taxonomy-based model for expertise extrapolation. In Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing, ICSC'10, pp 333-340, 2010.

[Caniza *et al.* 2014] Caniza H., A. E. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, G. Valentini, A. Paccanaro. GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology. Bioinformatics. Advance Access published, 2014.

[Cavy, 2015] Cavy B. Nouveaux anti-patrons pour l'identification des problèmes dans une ontologie. Rapport de projet de Recherche&Développement en cinquième année du département Informatique de Polytech Nantes, 2015.

[Cellier *et al.* 2010] Cellier P., T. Charnois, M. Plantevit. Sequential patterns to discover and characterise biological relations. In Computational Linguistics and Intelligent Text Processing, LNCS, pp 537–548, Springer, 2010.

[Choi *et al.* 2010] Choi S.S., Cha S.H., C. C. Tappert. A Survey of Binary Similarity and Distance Measures. Journal of Systemics, Cybernetics and Informatics, V. 8 No 1, pp 43-48, 2010.

[Champin&Solnon, 2003] Champin P-A, C. Solnon. Measuring the similarity of labeled graphs Proceeding ICCBR'03 Proceedings of the 5th international conference on Case-based reasoning: Research and Development, pp 80-95, 2003

[Charlet *et al.* 2000] Charlet J., M. Zacklad, D. Bourigault. Ingénierie des connaissances : évolutions récentes et nouveaux défis. Edition Eyrolles, 2000.

[Charlet *et al.* 2010] Charlet J., S. Szulman, N. Aussenac-Gilles, A. Nazarenko, N. Hernandez, N. Nadah, E. Sardet, J. Delahousse, V. Teguiak, and A. Baneyx. Dafoe : une plateforme pour construire des ontologies à partir de textes et de thésaurus. In 10ième Conférence Internationale Francophone sur l'Extraction et la Gestion des Connaissances, Hammamet, Tunisie, 2010.

[Chauvin, 2010] Chauvin L. Modèles de cartes cognitives étendues aux notions de contexte et d'échelle. Thèse en informatique de l'Université d'Angers, 2010

[Chauvin, 2012] Chauvin L. Analyse des résultats des mesures sémantiques. Rapport PostDoc Université de Nantes, 2012.

[Chulyadyo&Mittal, 2012] Chulyadyo R., N Mittal. Tools for ontology building from texts: Analysis and Improvement of the Results of Text2Onto, Project Rapport, DMKM, PolytechNantes, University of Nantes, 2012.

[Chulyadyo, 2012] Chulyadyo R. Improvement of Results of Text2Onto using a core ontology. Internship Report, DMKM, PolytechNantes, University of Nantes,2012.

[Chulyadyo *et al.* 2013] Chulyadyo R., M. Harzallah, G. Berio. Core Ontology Based Approach for Treating the Flatness of Automatically Built Ontology. In Proceeding of the 5th International Conferenceon Knowledge Engineering and Ontology development (KEOD 2013), 2013.

[Cicortas&Iordan, 2007] Cicortas, A., V. Iordan. Representing and Comparing Competences Using Agents. In Proc of the 4th International Symposium on Applied Computational Intelligence and Informatics, SACI '07, pp225 – 230, 2007.

[Cimiano&Volker, 2005] Cimiano, P., J. Volker. Text2onto-a framework for ontology learning and data-driven change discovery. In Proceedings of the 2nd Eur. SemanticWeb Conference, vol. 3513, édité par A. Montoyo, R. Munoz et E. Metais, pp 227–238, 2005.

[Cimiano *et al.* 2005] Cimiano P., A. Hotho, S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. In Journal of Artificial Intelligence Research (JAIR'05), V. 24, pp 305–339. AAAI Press, 2005.

[Collin, 2001] Collin R.. Gestion des connaissances et aide à la décision. Séminaire Connaissances, compétences et technologies. Pourquoi l'homme est la mesure de toute information ? . http://perso.wanadoo.fr/michel.grunstein/, 2001.

[Colucci *et al.* 2003a] Colucci S., T. Di Noia, E. Di Sciascio, F.M. Donini, M. Mongiello. Concept Abduction and Contraction in Description Logics. In *Proc. of the 16th Intl. Workshop on Description Logics (DL'03)*, V 81 of *CEUR Workshop Proceedings*, 2003.

[Colucci *et al.* 2003b] Colucci S, T. Di Noia, E. Di Sciascio, F. M. Donini, M.Mongiello, M. Mottola A formal approach to ontology-based semantic match of skills descriptions. Journal of Universal Computer Science, Special issue on Skills Management, 2003.

[Colucci *et al.* 2005] Colucci S., T. Di Noia, E. Di Sciascio, F.M. Donini, G. Piscitelli, S. Coppi. Knowledge Based Approach to Semantic Composition of Teams in an Organization. In *Proceedings of the 20th Annual ACM (SIGAPP) Symposium on Applied Computing SAC-05*, ACM, pp 1314-1319, 2005.

[Colucci *et al.* 2005a] Colucci, T. Di Noia, E. Di Sciascio, F.M. Donini, A. Ragone. Semantic-based automated composition of distributed learning objects for personalized e-learning. In *The Semantic Web: Research and Applications. 2nd European Semantic Web Conf.*, V. 3532, pp 633-648. 2005.

[Condamines, 2005] Condamines A., Sémantique et Corpus, Hermès Science Publications, ISBN 2-7462-1055-X, 2005.

[Corby *et al.* 2004] Corby O, Dieng-Kuntz R, Faron-Zucker C. Querying the Semantic Web with the NOYAUSE search engine. In (R. Lopez de Mantaras and L. Saitta eds) Proceedings of the 16th European Conference on Artificial Intelligence (ECAI'2004), subconference PAIS'2004, IOS Press, pp 705-709, 2004.

[Corcho *et al.* 2005] Corcho, O., M. Fernández-López, A. G. Pérez et A. López-Cima. 2005, Building legal ontologies with methontology and webode, Lecture Notes in Computer Science, V. 3369, pp 142–157.

[Cross *et al.* 2013] Cross V, X. Yu, X. Hu. Unifying ontological similarity measures: a theoretical and empirical investigation. Int J Approx Reason 54:861–75, 2013

[Culotta&Sorensen, 2004] Culotta, A., Sorensen, J. Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004.

[Damljanovic *et al.* 2009] Damljanovic D., F. Amardeilh, K. Bontcheva. CA Manager Framework: Creating Customised Work ows for Ontology Population and Semantic Annotation. The Fifth International Conference on Knowledge Capture (KCAP'09), 2009, Redondo Beach, California, United States.

[Darmont&Boussad, 2006] Darmont J., Boussad O.: Processing and Managing Complex Data for Decision Support. IDEA GROUP PUBLISHING, 2006.

[d'Aquin, 2012] d'Aquin M. Modularizing Ontologies. In M. C. Suarez-Figueroa, A. Gomez-Perez, E. Motta, & A. Gangemi, Eds.,Ontology Engineering in a Networked World, pp. 213–233. Springer Berlin Heidelberg, 2012.

[Dartigues, 2003] Dartigues C. Echange de données techniques dans un environnement coopératif. Thèse de l'université de Claude Bernard , 2003.

[Debruyne *et al.* 2013] Debruyne C., T.-K Tran, R. Meersman. Grounding ontologies with social processes and natural language. Journal on Data Semantics 2, 2-3, pp 89–118, 2013.

[Declerck *et al.* 2012] Declerck G., A. Baneyx, X. Aimé, J.Charlet. A quoi servent les ontologies fondationnelles ? 23èmes Journées francophones d'Ingénierie des Connaissances (IC 2012), pp. 67-82, 2012

[De Coi *et al.* 2007] De Coi L., E. Herder, A.W. Koesling, C. Lofi, D. Olmedilla, O. Papapetrou, W. Siberski. A Model for Competence Gap Analysis. WEBIST (3), pp 304-312 2007.

[Dehmer&Mowshowitz, 2011] Dehmer M., Mowshowitz A. A history of graph entropy measures. Information Sciences 181. pp57–78, 2011.

[Després&Szulman, 2007] Després, S., Szulman S. Merging of legal micro-ontologies from European Directives. *In Artificial Intelligence and Law*, 15, pp187–200, 2007.

[De Nicola&Missikoff, 2016] De Nicola A., M. Missikoff. A lightweight methodology for rapid ontology engineering. Communications in of the ACM Volume 59 Issue 3, pp 79-86, 2016.

[Di Jorio et al. 2007] Di-Jorio L., C Fiot, L Abrouk, D. Hérin, M. Teisseire. Enrichissement d'ontologie: Quand les motifs séquentiels labellisent des relations. BDA 2007.

[Diday&Kodratoff , 1991] Diday E., Y Kodratoff. Induction symbolique et numérique , Cépaduès Editions, 1991.

[Dieng *et al.* 1998] R. Dieng, O. Corby, A. Giboin, R. Brière. Methods and Tools for Corporate Knowledge Management. Projet ACACIA, Rapport de Recherche n°3485, 1998.

[Ding &Peng, 2004] Ding Z, Y. Peng. A probabilistic extension to ontology language owl. Proceedings of the 37th Annual Hawaii international Conference On System Sciences, 2004.

[Duque-Ramos et al. 2011] Duque-Ramos A., J.T. Fernandez-Breis, N. Aussenac-Gilles, R. Stevens. Oquare: Asquare based approach for evaluating the quality of ontologies. Journal of research and practice in information technology, 43, 159–173, 2011.

[Du *et al.* 2009] Du, Z., L. Li, C. F. Chen, P. S. Yu, J.W. Wang, G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. Nucleic Acids Research, 37(2), D345–D349, 2009.

[Elia&Margherita, 2015] Elia G., A. Margherita. Next-generation human resource management : A system for measuring and visualising professional competencies. In International Journal of Human Resources Development and Management 15(1):1, 2015.

[Ermine, 2000] Ermine J-L. Les systèmes de connaissances. 2ème édition revue et augmentée. Hermès Sciences Publications, 2000.

[Euzenat, 2008]. Euzenat J. Quelques pistes pour une distance entre ontologies. Actes 1er atelier EGC 2008 sur similarité sémantique, p. 51-66. 2008.

[Fahad *et al.* 2008] Fahad M., M.Abdul Qadir, S. A. H. Shah. Evaluation of ontologies and dl reasoners. In Zhongzhi Shi, E. Mercier-Laurent, and D. Leake, editors, Intelligent Information Processing IV, volume 288 of IFIP : The International Federation for Information Processing, Springer US, pp 17–27, 2008.

[Faure&Nédellec, 1998a] Faure, D., C. Nédellec. A corpus-based conceptual clustering method for verb frames and ontology acquisition. In LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications, edited by P. Velardi, pp 5–12, 1998.

[Faure&Nédellec 1998b] Faure, D., Nedellec C. ASIUM: learning subcategorisation frames and restrictions of selection. *In the 10th European Conference on MachineLearning, workshop on Text Mining*, 1998.

[Fayyad *et al.* 1996] Fayyad, U., Piatetsky-Shapiro G. , and P. Smyth., From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996.

[Feinerer&Hornik, 2014] I. Feinerer, K. Hornik. tm: A framework for text mining applications within R. R package version 0.6., 2014.

[Fensel *et al.* 1998] Fensel D., Harmelen F-V Wolfgang R. et Teije A-T., Formal support for Development of Knowledge-Based Systems, *Information Technology Management*, Vol. 2, No. 4, 1998.

[Fernandez *et al.* 1997] Fernandez, M., A. Gomez-Pérez, N. Juristo. Methontology : From ontological art towards ontological engineering. In Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering, pp 33–40, 1997

[Finkelstein *et al.* 2002] L. Finkelstein, E. Gabrilovich, Y. Matias, G. Wolfman E. Rivlin, Z. Solan, E. Ruppin. Placing search in context: The concept revisited. ACM Trans. Information Systems, 20(1), pp 116–131, 2002.

[Fortuna *et al.* 2007] Fortuna, B., M. Grobelnik, D. Mladenic.. Ontogen : Semi-automatic ontology editor. In Proceedings of the 2007 Conference on Human interface (HCI), pp 309–318, 2007.

[Fox *et al.* 1998] Fox, M.S., M. Barbuceanu, M. Gruninger, J. Lin. An Organisation Ontology for Enterprise Modeling. In Simulating Organizations: Computational Models of Institutions and Groups, M. Prietula, K. Carley & L. Gasser (Eds), Menlo Park CA: AAAI/MIT Press, pp 131-152, 1998.

[Fritzsche&Gruninger, 2016] Fritzsche D., Gruninger M. Ontologies within Semantic Interoperability Ecosystems. Ontology Summit 2016 Communique, 2016

[Ganascia, 1998] Ganascia J.G. Le Petit Trésor. Dictionnaire de l'informatique et des sciences de l'information. Édition Flammarion, 1998.

[Ganesan *et al.* 2003] Ganesan P., H. Garcia-Molina, J. Widom. Exploiting Hierarchical Domain Structure to Compute Similarity. ACM Transactions on Information Systems, V21 pp. 64-93, 2003.

[Gangemi&Borgo, 2004] Gangemi A. Borgo S. Eds. Workshop on Core Ontologies in Ontology Engineering, 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004), Whittlebury Hall, Northamptonshire,UK. CEUR online Proceedings, http://ceur-ws.org/Vol-118/ 2004.

[Gangemi *et al.* 2002] Gangemi A., N. Guarino, C. Masolo, A. Oltramari, L. Schneider. Sweetening ontologies with DOLCE. In 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2002), 2002.

[Gangemi *et al.* 2005] Gangemi, A., Catenacci, C., Ciaramita, M., & Lehmann, J. Ontology evaluation and validation: an integrated formal model for the quality diagnostic task. Rapport de travail, 2005. En ligne :*http://www. loa-cnr. it/Files/OntoEval4OntoDev_Final. Pdf.*

[Gangemi et al. 2006] Gangemi, A., Catenacci, C., Ciaramita, M., Lehmann, J.: Modelling ontology evaluation and validation. In Proc. of the 3rd. semantic web conf. (eswc2006). LNCS, no. 4011, pp 140–154, 2006.

[Gangemi, 2013] Gangemi A. A Comparison of Knowledge Extraction Tools for the Semantic Web. In: Cimiano P., Corcho O., Presutti V., Hollink L., Rudolph S. (eds) The Semantic Web: Semantics and Big Data. ESWC 2013. Lecture Notes in Computer Science, vol 7882. Springer, Berlin, Heidelberg, 2013.

[Gangemi&Presutti, 2009] Gangemi, A., V. Presutti. Ontology design patterns. In Handbook on Ontologies, edited by p R. Studer, S. Staab, 2e éd., International Handbooks on Information Systems, Springer, pp 221–243, 2009.

[Gaoussou *et al.* 2014] Gaoussou C., S. Despres, Moussa Lo. IDOSCHISTO : une extension de l'ontologie noyau des maladies infectieuses (IDO-Core) pour la schistosomiase. Catherine Faron-Zucker. IC- 25èmes Journées francophones d'Ingénierie des Connaissances, Mai, Clermont-Ferrand, France. pp.39-50, 2014.

[Garro&Palopoli, 2003] A. Garro, L. Palopoli. An XML MultiAgent System for e-Learning and Skill Management. In Agent Technologies, Infrastructures, Tools, and Applications for E-Services, LNAI 2592. Springer-Verlag, 2003.

[Gartner, 2002] Gartner T. Exponential and geometric kernels for graphs.In *NIPS*02 workshop on unreal data*, volume Principles of modeling nonvectorial data, 2002.

[Gaston *et al.* 2014] Mazandu G.K., N. J. Mulder. Information Content-Based Gene Ontology Functional Similarity Measures: Which One to Use for a Given Biological Data Type? 2014.

[Gaston *et al.* 2016] Mazandu G. K., E. R. Chimusa, N. J. Mulder. Briefings in Bioinformatics Advance Access published, 2016.

[Gayo, 2006] Gayo D. Une Architecture à base d'Ontologies pour la Gestion Unifiées des Donnéees Structurées et non Structurées. Interface homme-machine . Universitée Joseph-Fourier -Grenoble I, 2006.

[Gehlert&Esswein, 2007] Gehlert A., W. Esswein. Toward a formal research framework for ontological analyses, Advanced Engineering Informatics. Elsevier, 21(2), pp 119-131, 2007.

[Ghamnia, 2016] Ghamnia, A. Extraction de relations d'hyperonymie à partir deWikipédia. Actes de la conférence conjointe JEP-TALN-RECITAL 2016, V 3 : RECITAL, 2016.

[Gherasim *et al.* 2011a] Gherasim T., M. Harzallah, G. Berio, P. Kuntz. Analyse comparative de méthodologies et d'outils de construction automatique d'ontologies à partir de ressources textuelles. In Actes de la 11ème Conférence Francophone sur l'Extraction et la Gestion des Connaissances EGC'11. Cépaduès Edition, pp 377-388, 2011.

[Gherasim *et al.* 2011b] Gherasim T., M. Harzallah, G. Berio, P. Kuntz. Construction automatique d'ontologies : comparaisons expérimentales à différentes échelles. In acte de l'atelier ExCo'Co de IC'11. 2011.

[Gherasim *et al.* 2012] Gherasim T., G. Berio, M. Harzallah, P. Kuntz. Problems impacting the quality of automatically built ontologies. In Proceedings of the 8th Workshop on Knowledge Engineering and Software Engineering (KESE-2012), held in conjunction with ECAI, pp 25–32, 2012.

[Gherasim *et al.* 2013a] Gherasim T., G. Berio, M. Harzallah, P. Kuntz. Quality problem identification in automatically constructed ontologies. (Poster) in the 6th International Conference on Knowledge Capture (K-CAP'13), pp 137-138, Banff, Canada, 2013.

[Gherasim *et al.* 2013b] Gherasim T., M. Harzallah, G. Berio, P. Kuntz. Methods and tools for automatic construction of ontologies from textual resources: A framework for comparison and its application. In Advances in knowledge discovery and management, Springer, V. 471, pp.177–201, 2013.

[Gomez-Perez *et al.* 2001] Gomez-Perez A., M. Fernandez-Lopez, O. Corcho, Ontological engineering:With examples from the areas of knowledge management, e-commerce and the semantic web. Adv. Inf. And Know. Processing. Springer, 2001.

[Gonzalez&Dankel, 1993] Gonzalez, A. J., D. Dankel. The Engineering of Knowledge-Based Systems Theory and Practice. Prentice Hall, 1993.

[Gordon, 1981] Gordon A.D. Classification. Methods for the exploratory analysis of multivariate data. Chapman & Hall, 1981.

[Granadar, 2015] Granadar L. Evaluation of methods for taxonomic relation extraction from text. PhD thesis, Pontifícia Universidade Católica do Rio Grande do Sul, 2015.

[Green&Rosemann, 2000] Green, P. M. Rosemann. Integrated Process Modeling: An Ontological Evaluation. Information Systems. Elsevier, 25(2), pp. 73-87, 2000.

[Gruber, 1993] Gruber, T. A translation approach to portable ontology specifications. In Knowledge Acquisition, vol. 5(2), p. 199–220, 1993.

[Grenon et al. 2004] Grenon P, Smith B, Goldberg L Biodynamic Ontology: Applying BFO in the Biomedical Domain. In Ontologies in Medicine, 102: pp 20–38, 2004.

[Gruninger&Obrst, 2014] Gruninger M., L. Obrst, Semantic Web and Big Data meets Applied Ontology, the Ontology Summit. Applied ontology 9(2), pp 155-170, 2014.

[Gruninger&Fox, 1995] Gruninger, M., M.S. Fox,. Methodology for the design and evaluation of ontologies. Proceeding of the Workshop on Basic Ontological Issues in Knowledge Sharing, IJCAI, 1995.

[Guarino, 1997] Guarino, N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In M. T. Pazienza (ed.) Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology. Springer Verlag: 139-170, 1997.

[Guarino, 1998] Guarino, N. Formal ontology in information systems», dans Proceedings of the 1st International Conference on Formal Ontologies in Information Systems (FOIS 1998), Frontiers in Artificial Intelligence and Applications, vol. 46, IOS Press, pp 3–15, 1998.

[Guarino et al. 2009] Guarino, N., D. Oberle, S. Staab. What is an ontology ?. dans Handbook on Ontologies, édité par R. Studer et S. Staab, 2e éd., International Handbooks on Information Systems, Springer, pp. 1–17. 2009.

[Guarino&Welty, 2000] Guarino, N., C. Welty. A Formal Ontology of Properties. In, Dieng, R., O. Corby eds, Proceedings of EKAW-2000: The 12th International Conference on Knowledge Engineering and Knowledge Management. Springer LNCS V. 1937/2000. pp. 97-112, 2000.

[Guarinon&Welty, 2002] N. Guarino and C. Welty, Evaluating Ontological Decisions with OntoClean, In Communication of the ACM, 45(2), pp 61-65, 2002.

[Guerrero *et al.* 2015] Guerrero, D., L. Patricia, B. A. Lucia. Domain Analysis of the Research In Professional Competences, Technology and Engineering Cluster. Procedia-Social and Behavioral Sciences, 182, pp 163-172, 2015.

[Haidar-Ahmad *et al.* 2016] Haidar-Ahmad L., A. Zouaq, M.Gagnon. Automatic Extraction of Axioms from Wikipedia Using SPARQL. In: Sack H., Rizzo G., Steinmetz N., Mladenić D., Auer S., Lange C. (eds) The Semantic Web. ESWC 2016. Lecture Notes in Computer Science, Springer, vol 9989, pp 60-64, 2016.

[Harzallah, 2000] Harzallah M. Modélisation des aspects organisationnels et des compétences pour la réorganisation d'entreprises industrielles. Thèse de doctorat en automatique et productique au laboratoire de Génie Industriel et de Production mécanique, ENIM/Université de Metz, 2000.

[Harzallah *et al.* 2002] M. Harzallah, G. Berio, F. Vernadat. A formal model for assessing individual competence in enterprises. In Proceedings of the IEEE Conf. SMC02, Hammamet, Tunisie, 2002

[Harzallah&Vernadat, 2002] Harzallah M., F. B. Vernadat. IT-based competency modeling and management: from theory to practice in enterprise engineering and operations. Computers in Industry 48(2), pp157-179, 2003.

[Harzallah&Berio, 2004] Harzallah M., G. Berio. Competency Modeling and management: A case study (2004). 6ème internationale conférence on Enterprise Information Systems, (ICEIS'04), Editeurs: I. Seruca, J. Filipe, S. Hammoudi et J. Cordeiro , Publisher: University of Portucalense, pp 350-358, 2004.

[Harzallah, 2004] Harzallah M. Ontology of enterprise competencies. On the Workshop Enterprise Modelling and Ontology (EMO) of the 16th International Conference on Advanced Information Systems (CAISE'04), Springer Verlag, 2004.

[Harzallah *et al.* 2006] Harzallah M., G. Berio, F. Vernadat. Modelling and Analysis of individual competencies to improve industrial performances. IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, V 36, N 1, pp 187-207, 2006.

[Harzallah *et al.* 2007] Harzallah M., G. Berio, A.L. Opdahl. Incorporating IDEF3 into the Unified Enterprise Modelling Language (UEML), Atelier de EDOC Conference (EDOC '07), Eleventh International IEEE, Annapolis, USA, 2007.

[Harzallah *et al.* 2008] Harzallah M., G. Berio, E. Blanchard, P. Kuntz. Mesures sémantiques pour la comparaison des constructs des langages de modélisation d'entreprise.In Actes du 1er atelier Mesures de similarité sémantique, de EGC'08, Nice, France, 2008.

[Harzallah, 2010] Harzallah M. Tache 2.4 : Description et Avancement du développement des ontologies dans le projet ISTA3. Livrable intermédiaire de la tâche 2.4 de ISTA3, 2010.

[Harzallah, 2012] Harzallah M. Développement des ontologies pour l'interopérabilité des systèmes hétérogènes, applications aux cas industriels du projet ISTA3. Livrable final de la tâche 2.4 du projet ISTA3, 2012.

[Harzallah *et al.* 2012] Harzallah M., G. Berio, A. L. Opdahl. New Perspectives in Ontological Analysis: Guidelines and Rules for Incorporating Modelling Languages into UEML. Information Systems, Elsevier, V37, I5, pp 484-507, 2012.

[Harzallah *et al.* 2014] Harzallah M., G. Berio, G. Toader, P. Kuntz. Ontology Quality Problems – An experience with Automatically Generated Ontologies. In Proceeding of the 6th International Conference on Knowledge Engineering and Ontology development, 2014.

[Harzallah *et al.* 2015] Harzallah M., G. Berio, P. Kuntz. Towards an Approach for Configuring Ontology Validation. In Knowledge Discovery, Knowledge Engineering and Knowledge Management V 553 of the series Communications in Computer and Information Science, Springer, pp 388-404, 2015.

[Harzallah&Berio, 2015] Harzallah M., G. Berio. A unified framework for semantic comparison of objects: extension to semantic graph comparison, 19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES2015), 2015.

[Harzallah, 2016] Harzallah M. Anti-patrons partiels pour l'identification des problèmes de contradiction sociale dans une ontologie. Actes des 27ème journées francophones d'ingénierie des Connaissances (IC'2016), 2016.

[Harispe *et al.* 2014] Harispe S, D. Sánchez, S. Ranwez, S. Janaqi. J. Montmain. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. J Biomed Inform, 48, pp 38-53, 2014.

[Hartmann, 2004] Hartmann, J., P. Spyns, A.Giboin, D. Maynard, , R. Cuel, M. C. Suarez-Figueroa, Y. Sure. Methods for ontology evaluation. Tech. rept. Knowledge Web Deliverable D1.2.3, 2004.

[Hernandez, 2006] Hernandez N. Ontologies de domaine pour la modélisation du contexte en recherche d'Information. Thèse en Informatique de l'université de Toulouse 3, 2006.

[Hiermann&Hôfferer, 2003] Hiermann W, M. Hôfferer. A practical knowledge-based approach to skill management and personnel development, Journal of Universal Computer Science, V. 9 N. 12, 2003.

[Hitzler *et al.* 2016] Hitzler P., A. Gangemi, K. Janowicz, A. Krisnadhi, V. Presutti. Ontology Engineering with Ontology Design Patterns: Foundations and Applications. IOS Press, 2016.

[Hliaoutakis *et al.* 2006] Hliaoutakis A, G. Varelas, E. Voutsakis, EGM Petrakis, E. Milios. Information retrieval by semantic similarity. Int J Semant Web Inf Syst, 2, pp 55–73, 2006.

[Huang *et al.* 2014] Huang J., J. Dang, G. M. Borchert, K. Eilbeck, H. Zhang, M. Xiong, W. Jiang, H. Wu, J. A. Blake, D. A. Natale, M. Tan. OMIT: Dynamic, Semi-Automated Ontology Development for the microRNA Domain. PLOS ONE, 2014.

[Ibekwe-SanJuan, 2007] Ibekwe-SanJuan F. Fouille de textes : méthodes, outils et applications. Hermès-Lavoisier, 352p., Collection Systèmes d'information et organisations documentaires, 2007

[Issac *et al.* 2005] Isaac A., B. Bachimont, P. Laublet. Indexation de documents audiovisuels : ontologies, patrons de conception et d'utilisation. In 16ièmes Journées Francophones d'Ingénierie des Connaissances (IC'2005), 2005.

[Jiang&Conrath, 1997] Jiang J. J., D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. in *Proc. of int. conf. On Research in Computational Linguistics*. pp 19-33, 1997.

[Kashima *et al.* 2003] Kashima H., K. Tsuda, A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the* 20th *International Conference on Machine Learning (ICML)*, Washington, DC, United States, 2003.

[Kliegr, 2015] Kliegr T. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. In Web Semantics: Science, Services and Agents on the World Wide Web, V. 31, pp 59-69, 2015.

[Lame, 2002] Lame G. Construction d'ontologie à partir de texte, une ontologie du droit dédiée à la recherche d'information sur le Web, Thèse de doctorat, Ecole des Mines de Paris, 2002.

[Laublet *et al.* 2009] Laublet P., N. Aussenac-Gilles, V. Camps, P. Glize, N. Hernandez, H. Maurel, M. Mbarki, J. Mothe, B. Ralalason, A. Reymonet, B. Rothenburger, Z.Sellami, J.Thomas, A. Tissaoui. DYNAMO : DYNAMic Ontology for information retrieval Livrable Lot2, 2009.

[Laukkanen&Helin, 2005] Laukkanen M, H. Helin. Competence management within and between organizations. In proceeding of the CAISE'05 Workshops, Enterprise Modelling and Ontologies for Interoperability Workshop. V.2., pp 359-362, 2005.

[Le Boterf, 1997] Le Boterf G.. De la compétence à la navigation professionnelle. Les Editions d'organisation, 1997.

[LeCun, 2016] LeCun Y. Qu'est ce que l'Intelligence Artificielle. Collège de France. 2016 http://goo.gl/MpI1y4.

[Lee *et al.* 2008] Lee W-N, Shah N, Sundlass K, Musen M. Comparison of ontology-based semantic-similarity measures. AMIA Annu Symp Proc. pp 384–388, 2008.
[Lefever, 2016] Lefever E. A hybrid approach to domain-independent taxonomy learning. In Applied ontology 11(3), pp 255-278, 2016.
[Lenci&Benotto, 2012] Lenci A., G. Benotto. Identifying hypernyms in distributional semantic spaces. In SemEval. pp 75–79, 2012.
[Lenzerini, 2002] Lenzerini, M. Data integration: A theoretical perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. s.l.:ACM, pp 233-246, 2002.
[Leray, 2006] Leray P. Réseaux bayésiens : Apprentissage et diagnostic de systemes complexes. Modélisation et simulation. Habilitation à diriger la recherche en Informatique. Université de Rouen, 2006.
[Levy-Leboyer, 1996] Levy-Leboyer C. Evaluation du personnel : Quelles méthodes choisir? Les Editions, d'Organisation, Paris, France.1996.
[Lin, 1998] Lin D. An information-theoretic definition of similarity. In Proc. of the 15th Int. Conf. on Machine Learning Morgan Kaufmann, pp 296–304, 1998.
[Lindgren et al. 2003] Lindgren R., D. Stenmark, J. Ljungberg. Rethinking competence systems for knowledge-based organisations. *European Journal of Information Systems*, V.12, n. 1, pp18-29, 2003.
[Lopez *et al.* 2014] Lopez C., F. Segond, O. Hondermarck, P. Curtoni, L. Dini. Generating a resource for products and brandnames recognition. application to the cosmetic domain. In LREC, pp 2559–2564, 2004.
[Lord *et al.* 2003] Lord, P.W., R.D. Stevens, A. Brass, C.A. Goble. Semantic similarity measures tools for exploring the Gene Ontology. Pac Symp Biocomput, 601-612, 2003.
[Lucia&Lepsinger, 1999] Lucia, A. D., Lepsinger, R. The art and science of competency: Pinpointing critical success factors in organizations, Hardcover Edition, 1999.
[Kamel&Rothenburger, 2011] Kamel M., B. Rothenburger. Structures énumératives dans les textes vs. Structures hiérachiques dans les ontologies (regular paper). Dans : Journées Francophones sur les Ontologies (JFO 2011), Montréal. R. Bouazie, P. Bourque, G. Falquet (Eds.), ACM SIGGRAPH / IEEE, 2011.
[Kosmopoulos, 2010] Kosmopoulos A., E. Gaussier, G. Paliouras, and S. Aseervatham. The ECIR 2010 large scale hierarchical classification workshop. SIGIR Forum, 44, pp23–32, 2010.
[Krogstie et al. 1995] Krogstie, J., O. Lindland et G. Sindre. Defining quality aspects for conceptual models. In Proceedings of the IFIP8.1 Working Conference on Information Systems, Concepts : Towards a Consolidation of Views (ISCO3), pp 216–231, 1995.
[Kutz&Hois, 2012] Kutz O., J. Hois. Modularity in ontologies. In Applied Ontology 7 pp 109–112. IOS Press, 2012.
[Maedche&Staab, 2000] Maedche A. S. Staab. Discovering conceptual relations from text. In W.Horn, editor, Proceedings of the 14th European Conference on Artificial Intelligence (ECAI'00), IOS Press pp 321 – 325, 2000.
[Marreli, 1998] Marreli, A.F. An introduction to competency analysis and modeling, *Improvement, 37* (5), pp 8-17, 1998.
[Mathur&Dinakarpandian, 2012] Mathur S, D. Dinakarpandian. Finding disease similarity based on implicit semantic similarity. J Biomed Inform. 45(2), pp 363-71, 2012.
[Maynard *et al.* 2009] Maynard, D., A. Funk, W. Peters. Sprat : a tool for automatic semantic pattern-based ontology population. In Proceedings of the International Conference for Digital Libraries and the Semantic Web, pp 1–15, 2009.

[Mazandu&Mulder, 2012] Mazandu G.K., N.J. Mulder A topology-based metric for measuring term similarity in the Gene Ontology. Adv Bioinformatics, 17 pages, 2012.
[Mazandu&Mulder, 2013a] Mazandu GK, N.J. Mulder. DaGO-Fun: Tool for Gene Ontology-based functional analysis using term information content measures. BMC Bioinformatics, 2013.
[Mazandu&Mulder, 2013b] Mazandu, G. K., N. J. Mulder. Information contentbased Gene Ontology semantic similarity approaches: Toward a unified framework theory. BioMed Research International, 2013.
[Mazandu *et al.* 2016] Mazandu GK., E. R. Chimusa N. J. Mulder. Gene Ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Bioinformatics, 2016.
[Mazak &Huemer, 2015] Mazak A., C. Huemer. From business functions to control functions: Transforming REA to ISA-95. In Proceedings of the 17th IEEE Conference on Business Informatics (CBI 2015), D. Aveiro, U. Frank, K. J. Lin, and J. Tribolet, Eds., V. 1. pp 33–42, 2015.
[Mimno& McCallum, 2007] Mimno D., A. McCallum. Expertise modeling for matching papers with reviewers. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '07. ACM, pp 500-509, 2007.
[Mika, 2005] Mika P. Ontologies are us : A unified model of social networks and semantics. In Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen, editors, The Semantic Web - ISWC 2005, Proceedings of ISWC 2005, V 3729 of Lecture Notes in Computer Science, pp 522-536. Springer, 2005.
[Mikolov et al. 2013] Mikolov T, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean.. Distributed representations of words and phrases and their compositionality. In NIPS, pp. 3111–3119, 2013.
[Mille, 2001] A. Mille. Les connaissances : formaliser, raisonner, apprendre. Cours de DEA ECD, Lyon, 2001.
[Miller&Charles, 1991] Miller G.A., W.G. Charles. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6(1), pp 1–28, 1991.
[Mitchell *et al.* 2015] Mitchell T., W. Cohen, E. Hruschka, P. Talukdar, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis, T. Mohamed, N. Nakashole, E. Platanios, A. Ritter, M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta, X. Chen, A. Saparov, M. Greaves, J. Welling. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2015.
[Morineau *et al.* 2013] Morineau E., G. Nshimiye, V. Robidou. Enrichissement des patrons is-a dans Text2Onto. Rapport de Projet du Module Ontologie et web sémantique, PolytechNantes, Département Informatique, 2013.
[Monaghan *et al.* 2010] Monaghan F., G. Bordea, K. Samp, and P. Buitelaar. Exploring your research: Sprinkling some saffron on semantic web dog food. In Semantic Web Challenge at the International Semantic Web Conference, 2010.
[Muhlenbach&Lallich, 2010] Muhlenbach F., S. Lallich. Discovering research communities by clustering bibliographical data. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Main Conference Proceedings, pp 500–507, 2010.
[Mondary, 2011] Mondary T. Construction d'ontologies a partir de textes. L'apport de l'analyse de concepts formels. Université Paris-Nord - Paris XIII, 2011.

[Mousavi *et al.* 2014] Mousavi H., D. Kerr, M. Iseli, C. Zaniolo. Harvesting Domain Specific Ontologies from Text. In Proceedings of the IEEE International Conference on Semantic Computing , pp211-218, 2014.

[Mowshowitz&Mitsou, 2009] Mowshowitz A., Mitsou V. Entropy, Orbits, and Spectra of Graphs. In Analysis of Complex Networks Book: From Biology to Linguistics. Edited by Dehmer M., F. Emmert-Streib WILEY-VCH Verlag GmbH & Co. KGaA,Weinheim, 2009.

[Muhlenbach&Lallich, 2010] Muhlenbach F., S. Lallich. Discovering research communities by clustering bibliographical data. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Main Conference Proceedings, pp 500–507, 2010.

[Nazarenko *et al.* 2009] Nazarenko A., Zargayouna H., O. Hamon, J. Van Puymbrouck : Evaluation des outils terminologiques : enjeux, diffcultés et propositions. Traitement Automatique des Langues, 50(1), pp 257–281, 2009.

[Navigli&Velardi, 2004] Navigli, R., P. Velardi. Learning domain ontologies from document warehouses and dedicated web sites, Computational Linguistics. V. 30, no2, pp 151–179, 2004.

[Navigli&Velardi, 2005] Navigli, R., P. Velardi. Structural semantic interconnections : A knowledge based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 27, no 7, pp 1075–1086, 2005.

[Navigli *et al.* 2011] Navigli R., P. Velardi, S. Faralli. A Graph-based Algorithm for Inducing Lexical Taxonomies from Scratch. Proc. of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011), Barcelona, Spain, July 19-22th, 2011.

[Nebhi, 2013] Nebhi K. A rule-based relation extraction system using dbpedia and syntactic parsing. In Proceedings of the 2013th International Conference on NLP & DBpedia-Volume 1064, pp 74–79, 2013.

[Nédellec, 2007] Nédellec, C. Acquisition of relation extraction rules by machine learning. Rapport technique, Deliverable 6.4b for ALVIS (Superpeer semantic Search Engine) Project, 2007.

[Nédellec, 2013] Nédellec C. Extraction et modélisation de connaissance à partir de texte - Application à la biologie. Mémoire de HDR en Informatique, Université Blaise Pascal, Clermont Ferrand, 2013.

[Neuthaus&Vizedom, 2013] Neuthaus F., A Vizedom. Towards Ontology Evaluation across the Life Cycle. In ontolog.cim3.net/file/work/OntologySummit2013/.

[Nonaka&Takauchi, 1997] Nonaka I. & H. Takeuchi H. La connaissance créatrice : la dynamique de l'entreprise apprenante, De Boeck Université, Bruxelles, 1997,

[Novalija et al. 2011] Novalija I., D. Mladenić,. L. Bradeško. OntoPlus: text-driven ontology extension using ontology content, structure and co-occurrence information. In Journal of Knowledge-based Systems. V. 24 issue 8, pp 1261-1276, 2011.

[Ogata&Collier, 2004] Ogata, N. and Collier, N. Ontology Express: statistical and non-monotonic learning of domain ontologies from text. In Proc. Workshop on Ontology Learning and Population held at the 16th European Conference on Artificial Intelligence (ECAI'2004), pp. 43-48, 2004.

[OMG, 2011] OMG. Business Process Model and Notation. Version 2.0, www.omg.org/spec/BPMN/2.0/2011.

[Opdhal& Henderson-sellers, 2002] Opdhal, A.L., B. Henderson-sellers. Ontological Evaluation for the UEMLUsing the Bungue-Wander-Weber Model. Softeware and systems modeling. Vol. 1, no. 1, pp. 43-67, 2002.

[Opdahl *et al.* 2012] Opdahl A. L., G. Berio, M. Harzallah, R. Matulevičius. Ontology for Enterprise and Information Systems Modelling. in Journal of Applied Ontology, IOS Press. V. 7, no. 1, pp 49-92, 2012.

[Park *et al.* 2011] Park, J., W. Cho et S. Rho. Evaluating ontology extraction tools using a comprehensive evaluation framework. Data & Knowledge Engineering, vol. 69, p. 1043–1061, 2011.

[Pazienza *et al.* 2005] Pazienza, M. T., M. Pennacchiotti, F. M. Zanzotto. Terminology extraction: An analysis of linguistic and statistical approaches. In S. Sirmakessis (eds.), Knowledge mining: Proceedings of the NEMIS 2004 final conference, p. 255–279, Berlin Heidelberg. Springer, 2005.

[Panchenko *et al.* 2016] Panchenko A., S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S.P. Ponzetto, C. Biemann. Taxi: a Taxonomy Induction Method based on Lexico-Syntactic Patterns, Substrings and Focused Crawling. In Proceedings of the 10th International Workshop on Semantic Evaluation, 2016.

[Pantel *et al.* 2004] Pantel P., D. Ravichandran, E. H. Hovy. Towards terascale knowledge acquisition. In Proceedings of COLING-04. pp 771-777, 2004.[Pfeffer&Sutton, 2000] Pfeffer J., R.I. Sutton. The Knowing-Doing Gap: How smart companies turn knowledge into action, *Edition Hardcover*, 2000.

[Pearl, 1988] Pearl J. Probabilistic reasoning in intelligent systems: networks of plausible inference Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1988.

[Périnet&Hamon, 2014] Périnet A., Hamon T. Analyse et proposition de paramètres distributionnels adaptés aux corpus de spécialité. In Actes de JADT 2014 (Journées internationales d'Analyse statistique des Données Textuelles), pp 507-518, 2014.

[Petersen *et al.* 2016] Petersen N., I. Grangel-González, G. Coskun, S. Auer, M. Frommhold, S. Tramp, M. Lefrançois, A. Zimmermann: SCORVoc: Vocabulary-Based Information Integration and Exchange in Supply Networks. International Conference on Semantic Computing ICSC'16, pp 132-139, 2016.

[Poon&Domingos, 2010] Poon H., P. Domingos. Unsupervised Ontology Induction from Text. In the Proceedings ACL '10 of the 48th Annual Meeting of the Association for Computational Linguistics, pp 296-305, 2010.

[Posea, 2004] Posea V. Démarche de Construction d'une Ontologie des Compétences. Mémoire de Master ECD (Extraction des Connaissances à partir des Données), département Informatique, Université Polytechnique de Bucarest, Roumanie, 2004.

[Posea&Harzallah, 2004] Posea V., M. Harzallah 2004. Building an ontology of competencies, the workshop EMOI (Enterprise Modelling and Ontology: Ingredients for Interoperability) of the 5th International Conference on Practical Aspects of Knowledge Management (PAKM'4), publication series by Deutsche Bibliothek, 2004.

[Poveda *et al.* 2009] Poveda M., Suarez-Figueroa, M.C., Gomez-Perez, A. Common pitfalls in ontologydevelopment. pp 91–100 of: Proc. of the current topics in artificial intelligence (caepia'09), and 13th conf. On spanish association for artificial intelligence, 2009.

[Poveda *et al.* 2012] Poveda, M., M.C. Suarez-Figueroa, A. Gomez-Perez. Validating ontologies with oops! Knowledge engineering and knowledge management. LNCS. Springer pp 267–281, 2012.

[Punnarut&Srihar, 2010] Punnarut R., G. Sriharee. A researcher expertise search system using ontology-based data mining. In Conceptual Modelling 2010, Seventh Asia-Pacific Conference on Conceptual Modelling (APCCM 2010), pp 71-78, 2010.

[Patwardhan *et al.* 2003] Patwardhan S, Banerjee S, Pedersen T. Using measures of semantic relatedness for word sense disambiguation. In: Proc Fourth Int Conf Intell Text Process Comput; pp 241–57, 2003.

[Pesquita *et al.* 2008] Pesquita C., D. Faria, H. Bastos, A. EN Ferreira, A. O. Falcao, F.M. Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9(Suppl 5): S4, 2008.

[Pesquita *et al.* 2009] Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity inbiomedical ontologies. PLoS Comput Biol, pp 5-12, 2009.

[Petasis *et al.* 2013] Petasis G., R. Möller, V. Karkaletsis. BOEMIE: Reasoning-based Information Extraction. Dans Proceedings of the 1st Workshop on Natural Language Processing and Automated Reasoning co-located with 12th International Conference on Logic Programming and Nonmonotonic Reasoning. V. 1044, pp 60–75, 2013.

[Pedersen *et al.* 2004] Pedersen T., S. Patwardhan, J. Michelizzi. Wordnet similarity -measuring the relatedness of concepts. in Proc. 5th Ann. Meet. North American Chapter Assoc. Comp. Linguistics, pp 38–41, 2004.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. AI Magazine 11(5), pp 68-70, 1991.

[Pinto *et al.* 2004] Pinto, H. S., C. Tempich et S. Staab. Diligent. Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In Proce. of the 16th European Conference on Artificial Intelligence (ECAI 2004), pp. 393–397. 2004.

[Pirrò&Euzenat, 2010] Pirrò G., J. Euzenat, A feature and information theoretic framework for semantic similarity and relatedness, in: Proceedings of International SemanticWeb Conference, vol. 1, pp.615–630, 2010.

[Pomohaci, 2006]. Pomohaci E.. Development of an ontology diagnostic and validation tool, Rapport de Master ECD (Extraction des Connaissances à partir des Données), département Informatique, Université Polytechnique de Bucarest, Roumanie, 2006.

[Prié&Garlatti, 2004] Prié Y., S.Garlatti. Métadonnées et annotations dans le Web sémantique. In Le Web sémantique, Charlet J., Laublet P. & Reynaud C. (Ed.), Hors série de la Revue Information - Interaction -Intelligence (I3), 4(1), Cépaduès, 2004, pp 45-68, 2004.

[Probst *et al.* 2000] Probst G., S. Raub et K. Romhardt. Managing Knowledge : Building Blocks for Success. Chichester : John Willey & Sons, 2000.

[Quazi&Qadir,2011] Quazi N. I., M. A. Qadir. Algorithms for the evaluation of ontologies for extended error taxonomy and their application on large ontologies. J. UCS, 17(7) pp1005–1020, 2011.

[Rada *et al.* 1989] Rada, R. H. Mili, E. Bicknell, et M. Blettner. Development and application of a metric on semantic nets, IEEE Transactions on Systems, Man, and Cybernetics, 1, pp 17–30. 1989.

[Rauffet, 2014] Rauffet P., C. Da Cuna, A. Bernard. A dynamic methodology and associated tools to assess organizational capabilities. In Computers in Industry 65(1), pp158–174, 2014.

[Resnik, 1995] Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In Proc. O the 14th int. Joint conf. on Artificial Intelligence, 1, pp448–453. 1995.

[Resnik, 1999] Resnik P. Semantic similarity in a taxonomy : An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research.* pp 95–130, 1999.

[Richard *et al.* 2015] Richard M., X. Aimé, M.O. Krebs, J. Charlet. LOVMI : vers une méthode interactive pour la validation d'ontologies. In Actes des 26ème journées francophones d'Ingénierie desConnaissances (IC'2015). Plateforme AFIA'2015, 2015.

[Richards&Ogden, 1989] Richards IA., CK Ogden. The Meaning of Meaning. Harvest/HBJ 1989.

[Riesen&Bunke, 2010] Riesen K., H. Bunke. Graph Matching, in Graph Classification and Clustering Based on Vector Space Embedding. vol. 77, World Scientific, pp.15–34, 2010.

[Rodriguez&Egenhofer, 2003] Rodríguez M. A., M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering, 15, pp 442–456. 2003.

[Rohde, 2005] Rohde F. An Ontological Evaluation of Jackson's System Development Model, Australian Journal of Information Systems 2(2), pp. 77-87, 1995.

[Rohrer, 2012] Rohrer E. Making Ontology Relationships Explicit in a Ontology Network, The Semantic Web: Research and Applications, V.7295 of the series Lecture Notes in Computer Science pp 818-822, 2012.

[Roller *et al.* 2014] Roller S., K. Erk, G. Boleda. Inclusive yet selective: Supervised distributional hypernymy detection. In COLING: Technical Papers, pp 1025–1036, 2014.

[Roussey *et al.* 2009], Roussey C., O. Corcho, L.M. Blazquez. Vilches.: A catalogue of owl ontology antipatterns. Pages 205–206 of: Proc. of 5th int. conf. on knowledge capture, 2009.

[Roussey et al. 2010] Roussey C., F. Scharffe, O. Corcho, O. Zamazal. Une méthode de débogage d'ontologiesowl basées sur la détection d'anti-patrons. In Actes de la 21ème conférence d'Ingénierie des Connaissances, pp 43–54, 2010.

[Rubenstei&Goodenough, 1965] Rubenstein H., J.B. Goodenough. Contextual correlates of synonymy', Comm. ACM, 8(10), pp 627–633, 1965.

[Sampson&Fytros, 2008] Sampson D., D. Fytros. Competence Models in Technology-Enhanced Competence-Based Learning; In *Handbook on Information Technologies for Education and Training*, Heimo H., A. Kinshuk, J. M. Pawlowski, Demetrios Sampson (Eds.), pp 155-174. 2008.

[Sanchez *et al.* 2011] Sánchez D., M. Batet, D. Isern. Ontology-based information content computation. Knowl-based Syst 2011;24, pp 297–303, 2011.

[Sanchez *et al.* 2012] D. Sánchez1, M. Batet, D. Isern, A. Valls. Ontology-based semantic similarity: A new feature-based approach. Expert Systems with Applications: An International Journal, V. 39, I. 9, pp 7718-7728, 2012.

[Sánchez &Batet, 2011] Sánchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. Biomed Inform; pp49–59, 2011.

[Sanderson&Croft, 1999] Sanderson, M., B. Croft Deriving concept hierarchies from text. In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and developmentin information retrieval, ACM, pp 206–213, 1999.

[Schreiber *et al.* 1999] Schreiber G., H. Akkermans, A. Anjewierden, R. de Hoog, N.Shadbolt, W. Van de Velde, B. Wielinga. Knowledge Engineering and Management The CommonKADS Methodology, MIT Press, 1999.

[Schreiber *et al.* 2000] Schreiber G., H. Akkermans, A. Anjewierden., R. De Hoog, N. Shadbolt, W. Van de Velde, B.Wielinga. *Knowledge Engineering and Management*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, 2000.

[Schropp *et al.* 2013] G. Schropp, E. Lefever, V. Hoste. A combined pattern-based and distributional approach for automatic hypernym detection in Dutch. Proceedings of Recent Advances in Natural Language Processing. pp.593-600, 2013.

[Schut& O'Neil, 2013] R. Schutt et C. O'Neil. Doing Data Science : Straight Talk from the Frontline, O'Reilly Media, 2013.

[Seco *et al.* 2004] Seco N, T. Veale, J. Hayes. An intrinsic information content metric for semantic similarity in WordNet. In: López de Mántaras R, Saitta L, editors. 16th European conference on artificial intelligence, ECAI 2004, including prestigious applicants of intelligent systems, PAIS 2004, IOS Press; pp 1089–1090, 2004.

[Seitner et al., 2016] Seitner J., Bizer, C. Eckert, K. Faralli, S. Meusel, R. Paulheim, H. and Ponzetto, S. P. A large database of hypernymy relations extracted from the web. In proceeding of the Tenth international conference on Language Ressources and Evaluation (LREC2016), pp. 360-367, 2016.

[Sellami *et al.* 2011] Sellami Z, V. Camps, N. Aussenac-Gilles, S. Rougemaille. Ontology Co-construction with an Adaptive Multi-Agent System: Principles and Case-study. LNCS in Communications in Computer and Information Science, Ana Fred, Jan L. G. Dietz, Kecheng Liu, Joaquim Filipe (Eds.), Springer-Verlag, V. 128, pp 237-248, 2011.

[Sheena, 2016] Sheena N., M. J.Smitha , and J. Shelbi. Automatic extraction of hypernym and meronym relations in english sentences using dependency parser. In Procedia Computer Science, pages 539–546, 2016.

[Shvaiko&Euzenat, 2005] Shvaiko P., J. Euzenat. A survey of schema-based matching approaches. Journal on Data Semantics 3730, pp 146–171, 2005.

[Shvaiko&Euzenat, 2013] Shvaiko P. J. Euzenat. Ontology Matching: State of the Art and Future Challenges. IEEE Trans. Knowl. Data Eng. 25(1), pp158-176, 2013.

[Simperl&Tempich, 2009] Simperl, E., Tempich, C. Exploring the economical aspects of ontology engineering. Pages 445–462 of: Studer, R., & Staab, S. (eds), Handbook on ontologies, 2 edn. International Handbooks on Information Systems. Springer, 2009.

[Sitthisak *et al.* 2007] Sitthisak O., L. Gilbert, H.C. Davis, M. Gobbi. Adapting health care competencies to a formal competency model, *Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007)*, 2007.

[Smeulders *et al.* 2000] Smeulders A.W.M., M. Worring, S. Santini, A. Gupta, R. Jain. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), pp 1349–1380, 2000.

[Smith, 2007] Smith B. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. In Nat Biotechnol, pp 1251-1255, 2007.

[Snow *et al.* 2004] Snow R., D. Jurafsky, Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In NIPS, 2004.

[Stanford University, 2016] Stanford University, Artificial Intellience and LIFE IN 2030, 2016.

[Stankovic, 2010] Stankovic M., C. Wagner, J. Jovanovic, and P. Laublet. Looking for Experts ? What can Linked Data do for you ? In Proceedings of Linked Data on the Web 2010, LDOW 2010.

[Stankovic *et al.* 2011] Stankovic M., Jelena Jovanovic, and Philippe Laublet. Linked data metrics for flexible expert search on the open web. In Proceedings of the 8th extended semantic web conference on The semantic web : research and applications - Volume Part I, ESWC'11, pp 108-123, Heidelberg, Springer-Verlag, 2011.

[Staab *et al.* 2001] Staab, S., H. P. Schnurr, R. Studer et Y. Sure., Knowledge processes and ontologies, IEEE Intelligent Systems, vol. 16, no 1, pp. 26–34, 2001.

[Stumme, 2003] Stumme, G., M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic, N. Stojanovic, R. Studer, Y. Sure, R. Volz et V.

Zacharias. The Karlsruhe view on ontologies. Rapport technique, University of Karlsruhe, Institute AIFB, 2003.

[Suarez-Figueroa *et al.* 2008] Suarez-Figueroa, M., G. A. de Cea, C. Buil, K. Dellschaft, M. Fernandez-Lopez, A. Garcia, A. Gomez-Perez, G. Herrero, E. Montiel-Ponsoda, M. Sabou,B. Villazon-Terrazas, Z. Yufei. Neon methodology for building contextualized ontology networks. Rapport technique, Deliverable D3.3.2 for NeOn Project, http ://www.neon-project.org/nw/Deliverables, 2008.

[Suárez-Figueroa, 2010]. Suárez-Figueroa M. C. *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*. Thesis (Doctoral), Facultad de Informática (UPM), Espagne, 2010.

[Suárez-Figueroa *et al.* 2012] Suárez-Figueroa M. C., A Gómez-Pérez,. Fernández-López, The NeOn Methodology for Ontology Engineering M.C. Suarez-Figueroa et al. (eds.), Ontology Engineering in a Networked World, Springer-Verlag Berlin Heidelberg 2012.

[Suchanek *et al.* 2008] Suchanek F., G. Kasneci, and G.Weikum. YAGO - A Large Ontology fromWikipedia and WordNet. Elsevier Journal of Web Semantics, 6(3), pp.203–217, 2008.

[Suchanek, 2014] Suchanek F. Information Extraction for Ontology Learning. pp. 135-151.In Perspectives on Ontology Learning, Lehmann and Volker editors, Akademische Verlagsgesellschaft, 2014.

[Sure *et al.* 2000] Sure Y., A. Maedche., S. Staab. Leveraging Corporate Skill Knowledge: From ProPer to OntoProPer. In: Proceedings of the Third International Conference on Practical Aspects of Knowledge Management, Basel, Switzerland, 2000.

[Sussna, 1993] Sussna M. Word sense disambiguation for free-text indexing using a massive semantic network ; in *Proc. of the Second International Conference on Information and Knowledge Management*, poster, p. 67–74, 1993.

[Sussna, 1997] Sussna M. J. Text Retrieval Using Inference in Semantic, Metanetworks. PhD thesis, University of California, San Diego, 1997.

[Szathmary, 2006] L. Szathmary. Symbolic Data Mining Methods with the Coron Platform. PhD Thesis in Computer Science, Université Henri Poincaré – Nancy. 2006.

[Sy *et al.* 2012] Sy M-F, S Ranwez, J Montmain, A Regnault, M Crampes, V. Ranwez. User centered and ontology based information retrieval system for life sciences. BMC Bioinformatics ?13(Suppl 1):S4. 2012.

[Szulman, 2011] Szulman S. Une nouvelle version de l'outil Terminae de construction de ressources termino-ontologiques. IC 2011. 22èmes Journées francophones d'Ingénierie des Connaissances, 2011.

[Szulman, 2013] Szulman S. Méthode et outil de construction de RTO à partir de textes : Terminae. Rapport interne. https://www6.inra.fr/reseau-in.../Méthode+et+outil+de+construction+de+RTO.pdf, LIPN. Université Paris 13, 2013.

[Toussaint, 2011] Toussaint Y. Fouille de textes : des méthodes symboliques pour la construction d'ontologies et l'annotation sémantique guidée par les connaissances. Traitement du texte et du document. Mémoire de HDR en informatique, Université Henri Poincaré, Nancy, 2011.

[Tartir *et al.* 2010] Tartir, S., I.B. Arpinar, A.P. Sheth. Ontological evaluation and validation. In Theory and applications of ontology: Computer applications. Poli R., M. Healy, A. Kameas, (eds), Springer, pp 115–130, 2010.

[Tudorache *et al.* 2008] Tudorache T., Noy N., Vendetti J. Web-Protege: A Lightweight OWL Ontology Editor for the Web.Proceedings of the Fifth OWLED Workshop on OWL: Experiences and Directions, collocated with the 7th International Semantic Web Conference (ISWC-2008), 2008

[Tversky, 1977] Tversky A. Features of similarity, Psychol. Rev, 84, pp 327-352, 1977.
[Uschold *et al.* 1996] M. Uschold, M. King, S. Moralee, Y. Zorgios. The enterprise ontology. Technical Report AIAI-TR- 195, Articial Intelligence Applications Institute. The University ofEdinburgh, 1996.
[Uren *et al.* 2006] Uren V, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, F. Ciravegna. Semantic Annotation for Knowledge Management: Requirements and a Survey of the State of the Art. In Journal of Web Semantics: Science, Services and Agents on the World Wide Web (4), pp. 14-28, 2006.
[Uslu *et al.* 2017] Uslu T., W. Hemati, A. Mehler, D. Baumartz. TextImager as a Generic Interface to R. Conference Paper Jan 2017, Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017.
[Velardi *et al.* 2005] Velardi, P., R. Navigli, A. Cucchiarelli et F. Neri. Evaluation of ontolearn, a methodology for automatic learning of domain ontologies. In Ontology Learning from Text: Methods, Applications and Evaluation. P. Buitelaar, P. Cimiano et B. Magnini (eds), IOS. Press, V. 123, pp. 92–106, 2005.
[Velardi *et al.* 2013] Velardi, P., S.Faralli, R.Navigli. Ontolearn reloaded : A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3), pp 665–707, 2013.
[Vrandecic, 2009] Vrandecic, D. Ontology evaluation. In Handbook on ontologies, Studer, R., S. Staab, (eds), Springer, pp 293–314, 2009.
[Valtchev, 1999] Valtchev P. Construction automatique de taxonomies pour l'aide à la représentation de connaissances par objets. Thèse d'informatique, Univ. Grenoble 1, 1999.
[Vergnaud, 2003] Vergnaud N. De la Connaissance à la Compétence : Etude comparative des concepts Connaissance et Compétence. Mémoire de Master ECD (Extraction des Connaissances à partir des Données), Polytech Nantes, France, 2003.
[Vergnaud *et al.* 2004] Vergnaud N., M. Harzallah, H. Briand. Modèle de Gestion Intégrée des Compétences et Connaissances. Revue des nouvelles technologies de l'information, EGC'2004, D. Zighed et G. Venturini (Eds), Cépaduès Edition, pp. 159-170, 2004.
[Vernadat, 1999] F. Vernadat. Technique de Modélisation en Entreprise : Applications aux processus opérationnels. Economica, 1999.
[Von Krogh, 2000] G. Von Krogh. Enabling Knowledge Creation : How to Unlock the Mystery of Tacit Knowledge and Release the Power of Innovation. Oxford University Press, 2000.
[Volker et al. 2007] Völker J., Hitzler P., Cimiano P. Acquisition of OWL DL Axioms from Lexical Resources. In: Franconi E., Kifer M., May W. (eds) The Semantic Web: Research and Applications. ESWC 2007. Lecture Notes in Computer Science, V. 4519. Springer, 2007.
[Völker *et al.* 2008] Völker J., Fernandez Langa S., Sure Y. Supporting the Construction of Spanish Legal Ontologies with Text2Onto. In: Casanovas P., Sartor G., Casellas N., Rubino R. (eds) Computable Models of the Law. Lecture Notes in Computer Science, V. 4884. Springer, 2008.
[Wache *et al.* 2001] Wache H, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, S. Huebner. Ontology-based integration of information - a survey of existing approaches. In Proceedings of IJCAI workshop on Ontologies and Information Sharing, pp 108–117, 2001.
[Wand&Weber, 1988] Wand Y., R.Weber. An Ontological Analysis of some Fundamental Information Systems Concepts. In Proc. Ninth International Conference on Information Systems, DeGross, J. I., Olson, M. H. (Eds.), pp. 213–225, 1988.
[Wang *et al.* 2005] Wang H., M. Horridge, A. Rector, N. Drummond, J. Seidenberg. Debugging owl-dl ontologies: A heuristic approach. In Y Gil, E Motta, V.R Benjamins, M.A. Musen,

(edt), The Semantic Web, ISWC 2005, V3729 of Lecture Notes in Computer Science, Springer Berlin Heidelberg pp 745–757., 2005.
[Wang *et al.* 2007] Wang JZ, Z Du, R Payattakool, P.S. Yu, C.F. Chen. A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10), pp 1274–1281, 2007.
[Wang *et al.* 2010] Wang Y., W. Liu, D. Bell. A Structure-Based Similarity Spreading Approachfor Ontology Matching. Lec. Notes in Computer Science V. 6379, pp361-374, 2010.
[Wong *et al.* 2012] Wong W., W. Liu, M. Bennamoun. Ontology Learning from Text: A Look back and into the Future. ACM Computing Surveys (CSUR), 44(4), 20, 2012.
[Wu&Palmer, 1994] Wu Z., M. Palmer. Verb semantics and lexical selection. In Proc.32nd Annual Meeting Assoc. Computational Linguistics, pp. 133–138, 1994.
[Weiss *et al.* 2010] Weiss S M., N Indurkhya, T. Zhang, F. Damerau. Text mining: Predictive Methods for Analyzing Unstructured Information. Springer, 237 pages, 2010.
[Wielinga *et al*. 1992] Wielinga B., A. Schreiber, A. Breuker. KADS : a modelling approach to knowledge engineering. Knowledge Acquisition, 1992.
[Yang *et al.* 2016] Yang C., W. Shen, T. Lin, X. Wang. A hybrid framework for integrating multiple manufacturing clouds. The International Journal of Advanced Manufacturing Technology, V. 86, I. 1, pp 895–911, 2016.
[Yuancheng *et al.* 2010] Yuancheng T, N. Johri, D. Roth, J. Hockenmaier. Citation author topic model in expert search. In Proceedings of the 23rd International Conference on Computational Linguistics. pp 1265–1273, 2010.
[Zardi&Ben Romdhane, 2013] Zardi H., L. Ben Romdhane. An O(n2) algorithm for detecting communities of unbalanced sizes in large scale social networks. Knowl.-Based Syst. 37: pp.19-36, 2013.
[Zargayouna *et al.* 2016] Zargayouna H., C. Roussey, J.P. Chevallet. Recherche d'information sémantique: Etat des lieux. In Revue de Traitement Automatique des Langues. V. 56 N. 3, 2016.
[Zavitsanos *et al.* 2010] Zavitsanos E., G. Tsatsaronis, I. Varlamis, G. Scalable. Semantic Annotation of Text Using Lexical and Web Resources; Artificial Intelligence Theories Models and Applications, 2010.
[Zhang *et al.* 2006] Zhang P, Z. Jinghui, S. Huitao. Gene functional similarity search tool (GFSST). BMC Bioinformatics,7:135, 2006.
[Zouaq *et al.* 2011] A. Zouaq, D. Gasevic, M.Hatala. Towards open ontology learning and filtering. Information Systems, v.36 n.7, pp 1064-1081, 2011.
[Zouaq *et al.* 2012] Zouaq A., D. Gasevic, M. Hatala. Linguistic Patterns for Information Extraction in OntoCmaps. In Proceedings Of the 3rd Workshop on Ontology Patterns - WOP2012, in conjunction with the 11th International Semantic Web Conference, Boston, USA, 2012.
[Zhou *et al.* 2008] Zhou Z., Y Wang, J. Gu. A new model of information content for semantic similarity in WordNet. In: Yau SS, C Lee, Y.C Chung (eds). 2nd international conference on future generation communication and networking symposia, IEEE Computer Society, pp 85–90, 2008.

Contributions à l'Ingénierie des Connaissances : Construction et Validation d'Ontologie & Mesures Sémantiques

Résumé

L'ingénierie des connaissances (extraction, modélisation, capitalisation, exploitation…) a connu plusieurs mutations en s'adaptant au cours du temps à l'évolution des connaissances. Elle a notamment dû prendre en compte une évolution dans le temps des ressources des connaissances (experts, livres, bases de données, réseaux sociaux, tweeters, web des données…), de leurs formes (implicite, explicite, structurée, semi ou non structurée), de leurs utilisateurs (organisation, apprenant, utilisateur du web…), des supports de leur utilisation (livres, bases de données, systèmes à bases de connaissances, applications du web sémantique…), du volume et de la vitesse de multiplication de leurs ressources, des techniques de leur extraction, des langages de leur représentation… Dans ces différentes évolutions, l'ontologie a été reconnue comme une représentation sémantique pertinente des connaissances.

Je me suis intéressée depuis plus de 13 ans, aux problématiques liées aux ontologies, à leur construction, à leur validation et à leur exploitation, en ingénierie des connaissances. Mes contributions à ce domaine s'organisent autour de 3 axes. Le premier porte sur l'ingénierie des compétences et son articulation à l'ingénierie des connaissances, avec deux contributions majeures : (1) des modèles de connaissances (*i.e.* une ontologie noyau des compétences basée sur le modèle CRAI (Comptency Resource Aspect Individual) et le modèle CKIM (Competency and Knowledge Integrated Model) pour une représentation intégrée des compétences et connaissances) et (2) une architecture intégrante pour l'ingénierie des compétences. Cette architecture se base sur une modélisation ontologique et fine des compétences et permet de répertorier des techniques d'ingénierie des connaissances et les ressources associées pour l'extraction et la gestion des compétences. Elle a orienté mes travaux de recherche vers deux autres axes : l'axe 2 porte sur les méthodes et techniques d'ingénierie des connaissances pour la conceptualisation et la validation d'ontologie ; l'axe 3 porte sur les mesures sémantiques de comparaison d'objets, une mesure sémantique étant une technique de cette architecture, appliquée à une ontologie pour aider à accomplir certains processus d'une organisation.

Dans l'axe 2, j'ai proposé un cadre pour comparer des approches et outils de conceptualisation semi-automatique d'ontologie. J'ai développé deux approches de validation d'ontologie dans lesquelles j'ai cherché, d'une part, à coupler la validation à la conceptualisation, et d'autre part à leur intégrer des contraintes générées à partir d'une ontologie noyau formelle. La première approche est basée sur l'identification des problèmes pouvant nuire à la qualité d'une ontologie. La deuxième approche utilise des règles générées du méta-modèle d'annotation et/ou d'une ontologie noyau pour guider l'annotation d'un objet avec cette ontologie, enrichir cette ontologie et valider ces deux tâches. Ces deux approches pourraient se fusionner et s'étendre vers une approche semi-automatique de construction et validation intégrées d'ontologie, basée sur une ontologie noyau formelle.

Dans l'axe 3, j'ai proposé un cadre unifiant pour la définition de trois familles de mesures sémantiques de comparaison d'objets selon leur annotation par une ontologie : un objet peut être annoté par un concept unique, un ensemble de concepts ou un graphe sémantique d'une ontologie. En plus, ce cadre aide à analyser les résultats des mesures et à choisir une mesure adéquate pour une ontologie et une application données. Il se caractérise par l'utilisation d'une approche similaire pour l'approximation du contenu informationnel apporté par chaque type d'annotation, le contenu informationnel étant un élément principal et commun à la définition de ces trois familles de mesures..

Dans ce mémoire de HDR, je présente ces différentes contributions, en les positionnant par rapport à l'évolution des connaissances et par rapport à des travaux connexes de l'état de l'art. Je discute en conclusion la pertinence de mes travaux par rapport au challenge des masses de données et je présente mon projet de recherche et des perspectives liées à ce challenge.